eISSN 2587–8719

Философия

Журнал Высшей школы экономики

2025 — Т. 9, № 4

Philosophy

Journal of the Higher School of Economics

2025 · Volume 9 · № 4

National Research University Higher School of Economics
Faculty of Humanities · School of Philosophy and Culturology

Philosophy

2025 9 (4)

https://philosophy.hse.ru/ · philosophy.journal@hse.ru
eISSN: 2587–8719 · registration: ЭЛ № ФС 77–68963
Room 417a, 21/4 Staraya Basmannaya Str., 105066 Moscow, Russia · +7 (495) 772 95 90 * 12032

Национальный исследовательский университет «Высшая школа экономики»
Факультет гуманитарных наук · Школа философии и культурологии

Философия

2025 — Т. 9, № 4

https://philosophy.hse.ru/ · philosophy.journal@hse.ru
eISSN: 2587–8719 · регистрация: эл № ФС 77–68963
Старая Басманная 21/4, 105066 Москва (ком. 417а) · +7 (495) 7729590 * 12032

Contents

Practical Philosophy

Publications and Translations

Book Reviews

Содержание

Переводы и публикации

[Философская критика

From the Executive Editors of the Issue

Dear colleagues!

We are pleased to present the 4th issue of the 9th volume of "Philosophy. Journal of the Higher School of Economics." This special issue continues and expands the interdisciplinary dialogue initiated at the international discussion "Leveraging Artificial Intelligence to Enhance the Responsible Research and Innovation Framework" hosted by the School of Philosophy and Cultural Studies in March 2025. The event highlighted how the rapid development of AI challenges the RRI frameworks, and demonstrated that the ethics of technology requires more political philosophy. The contributions address new forms of responsibility, regimes of technocratic rationality, and the need for context-sensitive, politically grounded, and culturally plural approaches to responsible innovation in a multipolar world.

The article by Alexander Mikhailovsky and Elena Seredkina rethinks RI/RRI and AI ethics under conditions of a multipolar world, demonstrating the limits of their universalist claims. Drawing on the cases of China and Russia, the authors show that responsibility in the field of AI is shaped within different socio-cultural and political contexts. As an alternative, they propose the concept of a multipolar architecture of responsibility (MAR), oriented toward normative pluralism and dialogue between models of technological governance. The article by Dazhou Wang (China) substantiates the transition from engineering ethics, focused on individual moral judgment, to ethical engineering as a new discipline that embeds ethical principles into technological systems and governance processes. AI is examined simultaneously as an object of regulation and as a tool of ethical governance, enabling the operationalization of values and supporting proactive, scalable, and reflexive governance of emerging technologies. Armin Grunwald (Germany) demonstrates that digitalization and AI pose risks of gradual societal disruptions, including the erosion of freedom, responsibility, cognitive skills, and visions of the future. Grunwald argues for a reorientation of RI and technology assessment toward the early detection of such slow yet potentially destructive processes. Daria Bylieva and Alfred Nordmann (Germany) examine how AI undermines the conceptual foundations of RRI by acting as an "ontolytic" force — dissolving and reconstituting core categories like agency, authorship, and accountability that RRI traditionally seeks to govern. Elena Trufanova discusses trustworthiness

and responsibility as the key issues of the AI application and concludes that they cannot be delegated to artificial systems. While Peng Cheng and Zhihui Zhang (China) reframe responsibility from a phenomenological perspective — outlining governance mechanisms for Embodied AI — Elizaveta Karpova, finally, argues for a framework of distributed moral responsibility, which appears to better capture the hybrid and networked character of contemporary human-machine decision-making.

The "Practical Philosophy" section opens with an article by Andrey Shishkov describing the history of the development of object-oriented ontology and asserting its enduring legacy as a distinct and influential school of thought within post-continental philosophy. To learn about what "epistemic injustice" is, you can read the paper by Ekaterina Alekseeva. Through a study of medical contexts, it is demonstrated that overcoming epistemic injustice requires not only ethical correction of individual biases but also a more radical transformation of knowledge institutions to integrate diverse perspectives including "profane" knowledge. The reduction of knowledge to expert knowledge is not just an injustice; in the author's view, it is an epistemic misery that leads to a distorted picture of the world. The third paper in the section, Andrei Kravtsov's investigation of the cultural transfer as a complex process of semiotic adaptation focuses on the Meiji era (1868–1912) Japanese translation of Dostoevsky's *Crime and Punishment*. In his compelling study, Kravtsov examines how the Christian-existential themes of the source text undergo transformation through the lens of Buddhist-Confucian syncretism.

The "Publications and Translations" section features a conversation with Stelarc (Stelios Arcadiou), a Cyprus-born Australian performance artist known for radically merging the biological body with technology. In his introduction to this exclusive 2024 interview, Oleg Gurov analyzes his artistic practice and demonstrates how it challenges our fundamental assumptions about human identity and consciousness in an increasingly technological world.

The issue closes with a review by José Verissimo Teixeira da Mata (Brazil) of the second Russian edition of N. A. Vasiliev's *Imaginary Logic* (2025), which sheds light on the international impact of his work on non-classical logics.

Happy New Year and happy reading!

Alexander V. Mikhailovsky and Elena V. Seredkina

Искусственный интеллект и ответственные инновации в многополярном мире

Studies: Artificial Intelligence and Responsible Innovation in a Multipolar World

Mikhailovsky, A. V., and E. V. Seredkina. 2025. "Political Philosophy of Technology and Responsible Innovation in a Multipolar World: The Russian and Chinese Cases of AI Ethics" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 13–46.

Alexander Mikhailovsky and Elena Seredkina*

Political Philosophy of Technology and Responsible Innovation in a Multipolar World**

The Russian and Chinese Cases of AI Ethics

Submitted: Oct. 10, 2025. **Reviewed:** Nov. 11, 2025. **Accepted:** Nov. 16, 2025.

Abstract: This article introduces a political-philosophical framework for understanding Responsible Innovation / Responsible Research and Innovation and ethical AI governance within a multipolar world. It argues that although R(R)I is often presented as a neutral and universal model for aligning science and technology with ethical and societal values, it remains deeply embedded in Western liberal-democratic assumptions of deliberation, participation, and transparency. When viewed through non-liberal cultural and political traditions, these principles acquire new meanings, revealing the limits of normative universalism. Through a comparative analysis of Chinese and Russian approaches, this article challenges the Western liberal-democratic foundations of R(R)I and AI ethics. The Chinese model, rooted in Confucian harmony (*he*) and virtue (*de*), frames responsibility as moral mediation, while the Russian approach associates responsible innovation with contributing to the common good and technological sovereignty. The study critiques the asymmetric perception of both models of AI governance — where China's is seen as centralized yet harmonious, and Russia's as merely state-controlled — and offers a revised view of Russian "sovereign AI" as a collaboration framework enabling dialogue among government, industry, and science: the Russian AI Ethics Code reveals more interdisciplinary collaboration than typically acknowledged. Finally, the paper explores the notion of a multipolar architecture of responsibility to create space for cultural diversity within a shared humanistic vision. This framework positions science and technology as tools for global cooperation rather than geopolitical competition.

Keywords: Political Philosophy of Technology, Responsible Innovation, Responsible Research and Innovation, Multipolar World, Multipolar Architecture of Responsibility (MAR), Artificial Intelligence (AI), AI Ethics, Sovereign AI.

DOI: 10.17323/2587-8719-2025-4-13-46.

*Alexander Mikhailovsky, Doctor of Science in Philosophy; Associate Professor at HSE University (Moscow, Russia), amichailowski@hse.ru, ORCID: 0000–0001–9687–114X; Elena Seredkina, PhD in Philosophy; Associate Professor at Perm National Research Polytechnic University (Perm, Russia), elena_seredkina@pstu.ru, ORCID: 0000–0003–2506–2374.

INTRODUCTION

The concept of Responsible Innovation (RI), and its institutionalized form, Responsible Research and Innovation (RRI), is increasingly central to contemporary philosophy of technology and political philosophy of technology. Both concepts emerged within the framework of European science ethics to align technological progress with the fundamental societal values and have been widely adopted in science and technology studies (STS) and international science, technology, and innovation (STI) policies. There are some differences between RI and RRI, which can be summarized as follows (Srinivas, 2022: 31–32):

- RRI encompasses a broader spectrum, including science, innovation, and their impact on society, while RI focuses exclusively on the results of implementing technological innovations in society;
- Unlike RI, RRI places emphasis on the educational aspect, as well as the analysis of the relationship between science, technology, and society;
- RRI includes a wider context of STI policy, while RI is more oriented towards the industrial sector and innovation development.

In a narrower sense, RI can be understood primarily as a philosophical-normative approach that emphasizes anticipation of consequences, reflexivity, dialogical engagement, and ethical guidance in scientific and technological processes. In contrast, RRI is a formalized political-administrative strategy of the EU that incorporates mandatory components such as gender equality, public participation, ethics, open access, and science education in research and development projects. In other words, RRI represents a managerial and project-based mechanism embedded in state and international science funding systems.[1]

Both concepts are originally oriented toward the realities of Western liberal democracy, with its understanding of the subject as an autonomous individual, science as a public good, and innovation as socially accountable tools. It is precisely this cultural and political origin of RI that questions its universality in the context of global diversity in political regimes, cultural traditions, and philosophies of technology.[2] How can we deal with countries whose

[1] Between 2014 and 2020, the European Commission launched the "Science in and with Society" program, based on the general concept of RRI. In contrast, in the current "Horizon Europe" program, which runs from 2021 to 2028, the RRI policy concept is no longer mentioned as a cross-cutting theme and has almost completely disappeared from the underlying legal texts (Meier & Byland, 2020).

[2] In recent years, criticism of the RRI has intensified even in Europe. For example, some European scholars have examined why RRI has struggled to succeed as a concept for STI policy within the EU (Griessler et al., 2023).

political cultures don't share the liberal principles of individual rights, limited government, and the vision of science as an open marketplace of ideas? How should RI be applied in societies with a different model of interaction between the state, society, and science/technology? These questions become particularly relevant in the era of so-called multipolarity — not only geopolitical but also normative and value-based.

The emergence of RI was accompanied by hope for its universalization — the possibility of developing a unified normative framework for assessing and managing technological progress. However, on a global scale, it becomes clear that the universalist approach of RI faces significant challenges. It implies a certain ontology of the subject (the autonomous individual), a model of governance (democratic participation), as well as normative legitimacy through scientific rationality and consensus. The context of multipolarity, understood as encompassing normative and value-based diversity alongside geopolitical shifts, casts doubt on these very parameters. Different cultures have their own visions of the good, duty, and justice. In this sense, RI is not a neutral tool but a part of a broader cultural and political paradigm. For example, in the Confucian tradition of China, the values of harmony and stability dominate over individualism; in Russia, the idea of collective good and national sovereignty shapes the normative landscape differently than in EU countries.

In this regard, the concept of a multipolar architecture of responsibility (MAR) becomes crucial. It does not imply a rejection of RI but rather its reconceptualization as a framework open to intercultural dialogue and the pluralistic interaction of ethical and political rationalities. Such an architecture must take into account "the ethos of pluralization" (Connolly, 1995) — the differences in how moral subjects, collective goals, and relationships to technology are constituted in different societies.

A political-philosophical examination of RI should also involve questioning the very fundamental concepts of (1) innovation and (2) responsibility. The *policy framework of RRI* is based on the *discursive framework of RI*, which is derivative from the Green Agenda of the early 21st century. The European Commission prioritized research and innovation based on "grand challenges" like climate change and sustainable development. These challenges have been conceived as grand or global challenges and are usually called "wicked problems." However, these complex and rather abstract problems often exceed the competence of the diverse stakeholders (e. g., profit and non-profit organizations) who are supposed to solve these problems.

As Vincent Blok and Pieter Lemmens note, "because of these differences between various stakeholders, actual efforts to involve stakeholders in innovation processes are liable to failure" (Blok & Lemmens, 2015: 22). These fundamental differences and conflicts of interest create significant barriers to implementing RI in practice. Such challenges demonstrate that the idealized notion of seamlessly integrating ethical and societal considerations into innovation — thereby ensuring that scientific and technological progress aligns with societal needs — proves far more complex in reality than theoretical discussions often suggest.

To a certain extent, RI reawakens the political origins of innovation; and this is in striking contrast to what the presupposed concept of innovation suggests. If we are to think philosophically about Responsible Innovation, we have to ask what notion of innovation is a self-evident one in the literature on the theoretical frame of RI. The contemporary scholarship understands innovation as encompassing both technological development and its subsequent commercialization. This dual nature emerged with modern economic theories, which prioritized innovation's utilitarian value (Godin, 2009). RI discourse assumes the same economic rationality governing technological innovation. According to Blok and Lemmens (Blok & Lemmens, 2015), the economic paradigm underlying technological innovation — with its inherent growth imperative — may be fundamentally incompatible with RI. This very prioritization of profit and expansion has directly contributed to ecological degradation, resource exhaustion, and systemic global inequities, thereby undermining the ethical foundations of RI.

The most well-known operationalization of the concept of responsibility is the "care of the future through collective stewardship of science and innovation in the present" (Stilgoe et al., 2013: 1570). Inspired by the work of Hans Jonas (*Das Prinzip Verantwortung*, 1979), this view frames responsibility as a collective duty to ensure the long-term survival and flourishing of humanity on a planet with finite resources. Accordingly, Responsible Innovation moves beyond simply avoiding harm to actively shaping desirable and sustainable socio-technical futures through inclusive, anticipatory, and reflexive governance of the innovation process itself.

However, it appears that this proactive concept of responsibility — which must be steered by the macro-ethical imperative of sustainable development — conflicts with the economic rationality governing technological innovation. Now, we are witnessing how the Green Agenda is becoming obsolete and is being replaced by the agenda of Technological Sovereignty, which prioritizes the development of Artificial Intelligence infrastructure for

the new industrial politics. This shift raises a number of critical questions, in particular: What role can political philosophy play in strengthening the discursive framework of RI?

POLITICAL PHILOSOPHY OF TECHNOLOGY AND RESPONSIBLE INNOVATION IN A MULTIPOLAR WORLD

By introducing the concept of the political philosophy of technology for AI ethics, we rely on Carl Mitcham's thesis that ethics alone is insufficient to address the challenges posed by artificial intelligence and must be complemented by politics to cultivate human flourishing (Mitcham et al., 2024). The political philosophy of technology allows us to raise the question of technological and innovative sovereignty, since technological sovereignty co-emerged with nation-states that aim to promote specific technologies and engineering practices (Mitcham, 2022). What are the significant shortcomings of the ethics of technology, which fails to contend with global market dynamics and guide technological development within national political frameworks? One of the problems lies in the vision of market governance and innovation, which adheres to the principle of *technology neutrality*. This principle advocates for technology-neutral regulation, allowing market forces to determine the most successful solutions.

The existence of the problem at the European level is indirectly acknowledged by René von Schomberg, one of the architects of the RI concept and a former official of the European Commission:

> In Europe, the AI Act is hailed as a pioneering piece of legislation addressing the risks associated with innovative outcomes of AI technologies [...]. Yet it fails to account for the risk of becoming dependent on AI systems whose operational mechanics remain opaque (often protected under the property rights of private owners). While the Act ensures compliance with ethical standards — such as respecting individual autonomy — it falls short of defining socially desirable outcomes (Von Schomberg, 2025: 2–3).

Indeed, in light of cybersecurity risks, the issue of digital sovereignty has recently emerged as a pressing policy matter in both the EU and the USA. The political philosophy of technology could offer a way out of this situation. However, von Schomberg criticizes Mitcham's concept on the grounds that a movement toward technological sovereignty would mean a departure from an open economy and a "relatively open innovation ecosystem" (ibid.: 3). This objection stems from von Schomberg's view that technological sovereignty primarily implies control and is associated

with an ethics of moral constraint. It focuses on setting boundaries (e. g., "what we should not do") rather than articulating positive goals (e. g., "what we should do").

In contrast, we wish to leverage Mitcham's idea and therefore put forward two counterarguments demonstrating that RI needs to be enhanced by integrating the political philosophy of technology. First, the equation of technological sovereignty with state-controlled technology carries the implication that a nation-state must have access to the technological capabilities required to produce products domestically, rather than relying on global markets. This implication is flawed because it has not been proven that nation-states operate under conditions of "closeness" and require the same from engineering practices. Furthermore, pursuing technological development with political objectives does not necessarily entail any nationalistic implications that are incompatible not only with the European governance system but also with the emerging new architecture of a polycentric world.

Secondly, von Schomberg relies on RI, which "calls for a socio-political ambition not just to respect human agency but to enhance it through public investments in AI systems" (Von Schomberg, 2025: 3). Indeed, the ambition of RI transcends that of ethics. However, the crux is that the discursive framework of RI is grounded in the principles of deliberative democracy, fostering mutual responsiveness among societal actors and facilitating their collaboration in tackling shared challenges. Its core mechanism involves a structured, collaborative, and open process that reconciles the interests of innovators and various "stakeholders" and stimulates an inclusive public discourse concerning the societal implications of technological advancement to ensure that innovations are ethically sound, sustainable, and aligned with societal needs (Von Schomberg, 2013). While RI requires the participation of "stakeholders," including the public, in deliberative processes, a more fundamental question is overlooked, namely: Is the public admitted to the co-design and shaping of the very foundations of the RRI framework? (Penttilä, 2024). There is a certain risk that deliberation turns from a tool of democracy into a tool for legitimizing pre-made decisions, creating an appearance of participation while neutralizing real conflicts and inequalities.

These two arguments are sufficient to assert that the discursive framework of RI needs to be supplemented by political philosophy. At this point, we have to highlight a specific difficulty that this article aims to resolve. This difficulty concerns the normative horizon of RI and RRI, which must be called into question. RRI as a European policy framework aimed at integrating societal and ethical considerations into STI. It is commonly

described as "research and innovation conducted both for and with society." RRI aims to assist policymakers, researchers, businesses, and the public in better addressing the social and ethical implications of new technologies. Four dimensions of RRI are defined: anticipation, inclusivity, reflexivity, and responsiveness (Owen et al., 2012). Accordingly, the development of AI systems should be anticipatory, participatory, reflective, and responsive.

These principles stem from liberal democratic discussions on aligning research and innovation with societal demands, leading to the establishment of key overarching values — including anticipation, ethics, reflexivity, public and stakeholder engagement, openness, and mutual responsiveness. The scholarly lineage of RRI as inspired by STS reflects the co-constitution of science, technology, and society within the framework of European values (Von Schomberg, 2015). Given the context within which it has been developed, we can ask, how could RRI be transferred and practiced in other contexts that do not share similar founding values or norms?

The explicit focus on liberal democratic principles, norms, and methods like "engagement," "gender," "ethics," "science education," "open access," and "governance" makes an RRI agenda for such non-liberal countries like Russia and China highly problematic. In Russia, we are faced with the fact that society practically does not participate in decision-making related to scientific and technological development (Garbuk & Ugleva, 2024). The STS researchers in China suggest that public involvement in STI may clash with traditional Chinese norms, while existing platforms for civic participation remain inadequate. Main obstacles include conflicts with cultural values, undefined responsibilities, weak institutional frameworks for engagement, low public awareness of participation opportunities, and gaps in scientific intelligence (Zhao & Liao, 2019). These barriers appear to hinder the effective incorporation of societal values into innovation governance.

International scholars argue that the successful implementation of RRI in different socio-cultural contexts cannot be a simple "transfer" or "translation" of a ready-made European model. It must be a dialogical and transformational process of *transduction* that is responsive to local civic epistemologies and allows the very concept of RRI to be reconfigured and enriched through local engagement and practices (Doezema et al., 2019).

Lately, we have observed a growing number of publications on Technology Assessment and R(R)I in diverse contexts of Africa, China, Central and Eastern Europe, India, Japan, Latin America, and Russia (e. g., Grunwald, ed., 2024). While acknowledging the significant relevance of R(R)I for their countries, scholars are demonstrating a wide spectrum of critical approaches

and creating a discursive space. For instance, Krishna Ravi Srinivas proposes a hybrid and contextualized approach, which he argues is the most viable path forward for RRI in India; the scholar is convinced that "contextualizing RRI for India, particularly in the light of STI Policy (STIP)... and Scientific Social Responsibility (SSR) is feasible and desirable... RRI in theory and practice can benefit from interaction with ideas and practices developed in India" (Srinivas, 2022: 29). Poonam Pandey is more critical, arguing that the internationalization of RI has been hampered by processes of "othering." Specifically, this means imposing simplistic cultural frames (like "frugal innovation") and a "catching-up" narrative. Conversely, Indian actors "other" R(R)I itself, dismissing it as a European concept irrelevant to India's pressing developmental needs (Pandey, 2024). Comisso and co-authors, in turn, address a significant gap in the RRI literature: the under-development of a critical and self-reflexive pedagogy. The authors argue that as RRI expands beyond its European origins, its pedagogical practices must also be transformed to avoid reproducing epistemic hegemony and propose "critical resistance" as a theoretical framework (Comisso et al., 2024).

In light of this justified criticism, the "transduction" could remain a viable model, again, on the condition of political-philosophical reflection, which calls into question liberal normative universalism. Indeed, RRI can be formally separated from Western European values, and moreover, proceduralist democracy allows it to be viewed as a tool for realizing cultural differences. Darya Bylieva and Alfred Nordmann have highlighted a fundamental problem, arguing that there is one Western value that cannot be transferred: "RRI is expressive of a purely formal social framework that demands tolerance, if not indifference, towards different notions of the good life as if these notions could co-exist without contradiction" (Bylieva & Nordmann, 2025: 88). In this sense, the critical mission of R(R)I will remain unfulfilled unless it seriously takes into account not only socio-cultural specificities but also multiple visions of the good life.

In his recent chapter "Artificial Intelligence Liberalism" (Mitcham, 2024), Mitcham questions the ideology of (neo)liberalism and claims that the pressure of political challenges of AI is greater than what existing social-political theory can adequately address. Political philosophical criticism of AI demands questioning of this ideology because the AI community is locked in a symbiotic embrace with liberalism, the political theory that takes the freedom of equal individuals as the primary reality. This individualist emphasis is at the core of Euro-American political regimes, which, in the name of equality, resist admitting any hard and ethically or politically consequential

distinctions between the many and the few. The social ontology of the modern West prioritizes individual autonomy and freedom, even when trying to talk about the common good. The problem or paradox of liberalism is that in the name of freedom it imposes a kind of cosmopolitanism that makes it very difficult for those who may be committed to "alternative modernities" to the capitalist model of the West.

In other words, the relevance of a Western liberal-democratic framework — rooted in ideals of liberty, equality, and civic participation — remains uncertain in societies where notions of responsibility and inclusivity diverge. Pak Hang Wong (Wong, 2016) problematized the application of RRI seeded with liberal and democratic values into decent but non-liberal and non-democratic contexts, without acknowledging the impossibility of adopting RRI in such contexts. The Confucian turn toward technology has been systematically articulated in *Harmonious Technology* (Harmonious Technology..., 2021). The collection elaborates the ideas of harmony, ritual, self-cultivation of the engineer, and technological mediation as normative resources for an ethics of technology beyond Western individualism. This provides a philosophical legitimation for China's orientation toward social concord and moral order. Most recently, Wong has tried to diversify a keyword in the AI Ethics and Governance, proposing the Confucian "Trustworthy AI" (Wong, 2024).

Normative frameworks developed within the context of Western liberal democracy face limitations when implemented in different political and cultural systems. This necessitates a shift from normative universalism to a conceptual politics of difference — an approach that accounts for cultural specificity, forms of political legitimation, and historically established ethical regimes. Political philosophy of technology focuses on the differences in how moral subjects, collective goals, and engineering are constituted in different communities. It aims to reformat the RI discourse by including different normative traditions and expanding research on "non-Western" ethics in their relevance to RI. To achieve this, we make use of the concepts of MAR and Sovereign AI. These concepts are intended to substantiate the hypothesis that technological sovereignty is not reducible to an "ethics of moral constraint" and to demonstrate that, on the contrary, it could refer to a plurality of national models of engineering and AI technology development. Although these models indeed presuppose a considerable level of state control, they could at the same time establish positive goals, promote values and norms, and contribute to the worldwide AI ethics.

AI ETHICS IN A MULTIPOLAR WORLD: FROM GLOBAL RESPONSIBLE INNOVATION TO A MULTIPOLAR ARCHITECTURE OF RESPONSIBILITY

The first waves of global AI ethics sought to establish a universal set of guiding principles — dignity, justice, transparency, safety — and thereby to create a shared normative platform. However, a large-scale meta-review of more than 200 ethical documents revealed that, in practice, there emerged not a single canon but rather a family of overlapping yet culturally distinct sets of norms. As the authors note, the goal of their review was to determine whether there is a global consensus in AI ethics — a goal that ultimately proved elusive (Corrêa et al., 2023).

Another study analyzed more than sixty national strategies and AI programs developed by governments, international organizations, and corporate actors, including the EU, Canada, China, the OECD, UNESCO, the World Economic Forum, Google, IBM, and others (Roche et al., 2023). The researchers compared these documents with the frameworks of RRI and Value Sensitive Design to assess how effectively they incorporate values such as inclusivity, gender equality, cultural diversity, and the participation of vulnerable groups. Nearly 80% of the analyzed documents contained references to diversity and inclusion as ethical principles of AI, yet only about 20% provided concrete mechanisms for their implementation. This led the authors to describe a phenomenon of so-called "rhetorical inclusion," in which values are present in the language but absent in political practice. The authors further note that Western documents on AI ethics tend to formulate principles in universal terms — "human dignity," "justice," "accountability" — but these categories are not neutral; they reflect Western political and cultural norms grounded in individualism and liberal human rights. In other words, behind the universal language lies an asymmetry of voices, where the values and experiences of the Global South remain unheard.

Analyses of multilateral initiatives — including those of the UN, UNESCO, OECD, the Council of Europe, and the G20 — converge on the conclusion of normative fragmentation. Soft-law instruments and non-binding recommendations predominate, while progress toward a formal treaty is repeatedly obstructed by cultural, political, and sovereignty-related divides. A characteristic statement encapsulates this tension: "a treaty is needed, needed now," yet the path toward it remains blocked by diverging interests and contextual interpretations (González Peralta, 2022). Furthermore, González Peralta emphasizes that AI ethics norms are not universal in practice; they are interpreted and implemented through the prism of national and cultural

traditions. International organizations, therefore, are compelled to balance between universal human rights and the diversity of cultural values.

Recent literature identifies a broader shift toward normative multipolarity: different centers of power now embed AI within their own ideological and political paradigms of governance. In particular, for China, AI is described as being embedded within a governance model focused on centralized control, ideological alignment, and social stability (Papadopoulou, 2025). In this sense, AI becomes a carrier of harmony, stability, sovereignty, and state leadership — an alternative to the liberal-deliberative model that dominates Western discourse.

The initial universalism of AI ethics has neither been empirically confirmed nor institutionally realized. The cumulative analysis of ethical and regulatory documents demonstrates a marked cultural variability and a persistent gap between declared principles and actual practices. Multilateral efforts to develop global ethical frameworks have encountered the barriers of national sovereignty and geopolitical interest, while national strategies have articulated their own culturally conditioned modes of the good and responsibility. Under these circumstances, the further development of AI ethics requires a transition toward a multipolar architecture of responsibility — one based on minimal universal orientations, mechanisms of cross-cultural translation of values, and distributed institutions capable of maintaining equilibrium among multiple centers of power.

THE CHINESE CASE: ETHICS OF AI AND THE LOGIC OF SOCIAL HARMONY

R(R)I is often presented as a neutral "ethics-as-procedure." In practice, however, it is embedded within concrete cultural and political ontologies of the good, duty, and justice. Within MAR, the starting point is the recognition of normative pluralism: different societies institutionalize responsibility in different ways because they hold different conceptions of the human being, authority, the common good, and the role of technology. This framework does not reject R(R)I; rather, it reconfigures it in accordance with local axioms, forms of legitimacy, and historically shaped regimes of ethical reasoning.

This Confucian orientation is evident already at the level of national AI governance principles. In 2021, the Ministry of Science and Technology (MOST) and the National Committee for Artificial Intelligence Management published "Ethical Norms for Next-Generation Artificial Intelligence," which listed "harmony and friendliness" as one of the main ethical imperatives, along with overall responsibility and security/control (Ethical Norms

for New Generation..., 2021). These principles define trustworthy AI primarily as an instrument of social coherence and stability, rather than merely as a guarantor of individual rights.

In the Chinese context, ethical principles do not remain mere soft law: by 2021, specific provisions of these ethical norms were integrated into the country's broader and stricter regulatory frameworks for data and digital platforms — such as the Personal Information Protection Law (PIPL) and the Data Security Law (DSL). Scholars describe this as a tighter model of digital transformation governance, consistent with the Confucian hierarchy where duty outweighs rights and the group outweighs the individual. It represents a form of centralized (party-state) coordination, supported by hierarchically organized "networked participation" of trusted corporations and academic institutions (Qiao-Franco & Zhu, 2022).

THE PHENOMENON OF THE LITERATI IN CHINESE CULTURE: THE MORAL-POLITICAL STRUCTURE OF RESPONSIBILITY

To understand why the principle of public participation cannot be transferred into the Chinese context, one must turn to the historical figure of the *literati* (Mitcham et al., 2024: 7–8). The Chinese literati were not merely scholars or intellectuals in the Western sense; they embodied a unique synthesis of moral self-cultivation, public service, and political responsibility. Emerging from the Confucian order of governance, the literati served as custodians of both ethical norms and political legitimacy. The ideal of the *junzi* ("gentleman," "superior person"), uniting inner moral perfection with social harmony, lies at the heart of this intellectual tradition.

Within this paradigm, knowledge does not have a liberating or oppositional character but is ritually interwoven into the fabric of governance: wisdom acquires meaning only insofar as it contributes to the moral cultivation of the ruler and the stability of the community. This Confucian model produced a technopolitical continuity that continues to shape the Chinese understanding of STI. In the absence of deliberative democratic institutions, the moral-intellectual elite functions as a mediator between the state and society. The literati act as epistemic and ethical translators of state objectives, ensuring that scientific and technological progress remains aligned with the moral ideal of harmony and collective flourishing. Thus, public participation in the Western sense — as horizontal and open discussion — is replaced in China by a culture of virtuous mediation, in which scholars and experts act as the moral representatives of the people rather than facilitators of dialogue (Wang & Long, 2023; Zhao & Liao, 2019).

In contemporary China, the figure of the *digital literati* represents not a rupture but a transformation of the classical literati tradition. Whereas historical *junzi* legitimized authority through moral teaching and Confucian education, the digital literati of the twenty-first century perform a comparable function within the architecture of technopolitical governance. They mediate between scientific knowledge, ethical reflection, and state policy, forming an epistemic elite that defines what responsible innovation means in the Chinese context. These digital literati are not merely engineers or bureaucrats, nor simply academic philosophers. They constitute a hybrid class of moral-technical translators — scientists, researchers, and policymakers whose competence integrates data science, ethics, and governance.

This techno-moral orientation expresses a transformation of the Confucian ethics of junzi into a form of algorithmic politics of virtue. The digital literati legitimize the governance of AI and data not through public debate but through ritualized expertise: expert consensus, official "white papers," and government committees function as the modern analogues of the imperial examination system, filtering both moral and technical competence. Thus, the digital literati occupy the same symbolic space as their predecessors — as mediators between the moral way (*dao*) and the technical knowledge (*zhi*).

From the MAR perspective, this configuration reveals how China creates its own form of RI — not by copying Western models of transparency and participation, but by embedding technology within a moral-bureaucratic cosmology. Here, responsibility is understood through the categories of virtue (*de*) and harmony (*he*) rather than through procedural inclusivity. In this way, the digital literati replace public participation with moral mediation, constructing a model of ethical governance without deliberative democracy — yet one that remains internally coherent within China's cultural logic of legitimacy.

CHINESE LARGE LANGUAGE MODELS AND THE ETHICS OF HARMONY

Contemporary Chinese models of AI also reveal a profound connection with the cultural foundations of Chinese society. Unlike Western large language models (LLMs), whose architectures and training data are oriented toward universalist principles of autonomy and rational transparency, Chinese LLMs embody traditional values of social harmony, collective consensus, and moral duty.

The study by Wu and co-authors (2025), which introduces the Chinese Value Corpus — a large-scale dataset of ethical and value-based rules for aligning LLMs — demonstrates that among the core values embedded

within Chinese AI architectures, the central position is occupied by harmony, defined as "social stability, class harmony, and the harmonious development of human beings and nature." This indicates that the cultural orientation toward harmony, deeply rooted in the Confucian tradition, has become part of the normative architecture of Chinese AI, guiding its outputs toward the preservation of consensus and social stability (Wu et al., 2025).

Empirical findings further confirm that Chinese LLMs are trained not only on linguistic material but also reflect the high-context, collectivist communication culture characteristic of China. In a comparative study, Liu and co-authors (2025) observe that Chinese models such as DeepSeek and Qwen tend to avoid direct confrontation employ softened and polite forms of disagreement, displaying a deferential attitude and respect for hierarchy. Unlike Western systems, they operate within the logic of the "preservation of face" (*mianzi*) and the "maintenance of social harmony." As the authors emphasize, "overall, Western LLMs reflected low-context, individualist norms, while Chinese LLMs embodied high-context, collectivist etiquette" (Liu et al., 2025: 4).

These results suggest that Chinese artificial intelligence functions as a kind of cultural agent that transmits the values of the society that created it. Far from being a neutral instrument, it reproduces the ethical and cultural matrix of Chinese civilization, in which social harmony, collective responsibility, and respect for hierarchy are regarded as the key organizing principles of social order. The Chinese trajectory of LLM development demonstrates that AI does not necessarily reflect universal behavioral norms or moral orientations. Instead, it can embody locally grounded ethical frameworks, confirming the MAR logic, where AI systems become mirrors of their respective moral worlds rather than vehicles of global uniformity.

COMPARISON BETWEEN THE CHINESE AND WESTERN EUROPEAN MODELS OF AI ETHICS IN THE CONTEXT OF MULTIPOLAR ARCHITECTURE OF RESPONSIBILITY

In Western European approaches to AI ethics, as M. Coeckelbergh argues, the central category is the common good, understood in the sense of republican political philosophy — as that which must be the object of democratic deliberation, collective action, and civic virtue (Coeckelbergh, 2024). The common good is not predetermined; it emerges through the procedure of deliberation, that is, through a public process in which citizens, experts, and policymakers collectively determine how technologies ought to serve society.

The good does not exist independently of social dialogue — it is continuously co-defined and renegotiated through public reasoning (Coeckelbergh, 2024).

In the Chinese model, the concept of the good has a different metaphysical and normative foundation. Here, not deliberation but harmony represents the highest form of the good. In the Confucian tradition, goodness is not defined through debate or consensus but through social concord and stability, which express the ideal of harmony. It is not the outcome of negotiation but an ontological condition of proper order, maintained through morally virtuous governance (*dezheng*). Within this logic, the good is not an object of agreement but a goal of moral cultivation and ethical guidance for society.

Thus, whereas the Western European model is oriented toward a procedural conception of the good, the Chinese model is oriented toward state — harmonious condition of moral and social order and Chinese AI ethics institutionalizes harmony and stability as supreme regulative principles. Within this framework, the notion of RRI is redefined: participation and trust are not conceived as open, agonistic deliberation but as hierarchically moderated attunement among actors around the maintenance of social order.

THE RUSSIAN CASE: AI ETHICS AS A LOCALIZED MODEL OF RESPONSIBLE INNOVATION

The Russian model of AI ethics is an important example of the localized development of the concept of RI, adapted to a different political and cultural reality. Its formation occurs in the context of a strong institutional role of the state and a cultural orientation toward collective values and historical continuity. The strategic importance of AI implementation is closely linked to Russia's demographic and geographical challenges. In the context of population decline and labor shortages, particularly acute in remote and inaccessible regions, automation is increasingly seen as a viable solution to maintain productivity in both industry and social services. These issues are especially relevant in the context of government initiatives aimed at developing new territorial-industrial zones and smart cities in Siberia, the Far East, and the Arctic North.

One striking example is the official adoption of the "Arctic 2035" strategy, which foresees the deployment of autonomous production systems controlled by AI in the Far North (Ukaz..., 2020). Within this strategy, robots and robotic systems will compensate for labor shortages and help exploit resource-rich, sparsely populated areas. The intellectualization and use of robots will not only reduce the labor deficit but also increase labor productivity. Creat-

ing a wide range of scalable low-population industries, and consequently boosting the country's GDP, will serve as a source of psychological uplift for Russians, as has occurred in Russian history (Vasil'yev, 2022: 55–56).

The centralized nature of AI development management in Russia is another distinctive feature of its national strategy. Unlike Western models that emphasize public-private partnerships and delegate a significant part of the innovation process to industrial enterprises (Kamolov et al., 2022; Ulnicane, 2021), the Russian approach keeps decision-making at the federal level. The initiative for the extension, implementation, and control of the strategy rests directly with the President of the Russian Federation, highlighting the guiding and supervisory role of the state in the national AI agenda. This centralization is accompanied by a strategic focus on the development of education and human capital. Special attention is given to supporting fundamental Russian traditions in mathematics and natural sciences, many of which were established during the Soviet period. Accordingly, the strategy fosters the development of domestic expertise and training of qualified specialists to ensure the autonomous development of AI technologies (Kamolov et al., 2022).

In terms of ethical and applied AI regulation, Russian theory and practice are largely oriented towards international standards.

> Ethical issues related to AI are faced by the entire world community, which means that it is necessary to develop some normative document that all countries can follow to formulate specific standards or recommendations that take into account the values, cultural traditions, and moral norms of different countries (Leushina & Karpov, 2022: 125).

As a model for the general framework, "Ethics of Artificial Intelligence: The Recommendation," adopted by the UNESCO General Conference in 2021, is considered (UNESCO, 2021).

Philosophical and axiological considerations are also reflected in the Russian AI Ethics Code (Kodeks..., 2021). The document directly mentions respect for cultural and linguistic diversity, the preservation of national identity, and attention to the traditions of different peoples and social groups (§ 1.1). It emphasizes the importance of predictive research and ethical forecasting when implementing intelligent technologies into society. As stated in the Code:

> Making decisions in the field of AI use that significantly affect society and the state should be accompanied by a scientifically verified, interdisciplinary forecast of

> socio-economic consequences and risks and examination of possible changes in the paradigm of value and cultural development of the society (§ 2.1).

The Code highlights that AI systems do not have legal status or moral autonomy, and all responsibility for their functioning and the consequences of their use rests with humans. The Code requires risk assessment and analysis of the possible humanitarian consequences of AI at all stages of its lifecycle, as well as calls for precautionary measures and monitoring of negative outcomes in the short, medium, and long term.

> Ethical principles in the field of AI and their codification lay the foundation for a more detailed dialogue among AI participants, defining priorities and general rules in the absence of large-scale legislative regulation. They can be practically applied as solutions to ethical dilemmas and, to some extent, integrated into engineering and technical decisions (Maslova et al., 2022: 79).

Ethical AI regulation in Russia is a result of interaction across multiple levels — legal, expert, institutional, and cultural. The Russian model is characterized by a combination of "soft law" and strategic planning within a strong state vertical. The Russian regulatory framework treats ethics instrumentally, as a means to coordinate and oversee AI projects in socially sensitive areas like education, healthcare, and public administration. Embedded within the concept of technological sovereignty, ethics acts as both a protective barrier against external standards and a tool for legitimizing domestic AI reforms (Repin & Ignatyev, 2024). This creates a distinct state-expert coordination architecture, reliant on specialized knowledge and administrative resources rather than public participation (Maslova et al., 2022).

Thus, the political-legal framework of AI ethics in Russia is a system of institutional risk mitigation, where ethics functions not as an ideological postulate but as a managerial and regulatory tool aimed at balancing technological development with social stability.

TECHNOLOGY AS A FORM OF CULTURE: THE RUSSIAN AXIOLOGY OF RESPONSIBILITY

Compared with the UNESCO Recommendation on the Ethics of Artificial Intelligence and other similar documents issued by international organizations, the Russian Code is remarkably concise in its ethical and axiological part. The excerpts cited above — though undoubtedly relevant and well-founded — remain largely declarative in form and require further conceptual development and substantive elaboration. The following section

outlines the core ideas that may guide its future evolution and practical implementation.

The roots of the Russian philosophy of technology can be traced back to P. K. Engelmeyer (1855–1942), who was among the first to interpret technology as a cultural phenomenon rather than a merely material or instrumental process (Engel'meyyer, 1898). He argued that the technical sphere is inseparable from the moral and intellectual development of humanity and that philosophy must explore the role of technology as a cultural factor and a form of human creativity (Engel'meyyer, 2013).

V. M. Rozin, the Russian philosopher of technology, also argues that technology is not merely a material system but a form-generating element of culture, carrying within itself its value codes, anthropological presuppositions, and symbolic meanings. "As an event of culture, technology must correspond to its meanings and develops according to its inner forces" (Rozin, 2004). Therefore, the analysis of technical artifacts requires not only engineering but also cultural — hermeneutic interpretation, revealing which values and goals are realized through them.

Developing Rozin's idea of technology as a form of culture, one can argue that technical objects and systems never exist in a "pure" form and cannot be understood outside the cultural context in which they emerge. Technology always expresses a particular vision of humanity, society, and nature, rooted in the values of a community. It embodies not only material but also symbolic relations between the human being and the world, manifesting a cultural understanding of how one ought to act and what is considered right, useful, and harmonious. Consequently, the study of technical artifacts and technologies requires not only engineering or ethical evaluation but also worldview attitudes, uncovering which ideals and goals are materialized within them. In this sense, every culture generates its own type of projective thinking and a distinctive logic of technical rationality that reflects its worldview foundations.

Based on the philosophical and socio-political premises stated during the analysis of the Russian case, three ethical and cultural directions can be identified that could enrich the Russian Code of Ethics in the field of artificial intelligence:

- *Collective Good* — the priority of public and universal interests over private ones; understanding AI as a means of strengthening social solidarity rather than as a source of competition or division.

- ◇ *Cultural Identity* — recognition that technologies should serve the preservation and development of national culture, language, traditions, and spiritual foundations, rather than their erosion through globalization.
- ◇ *Technological Sovereignty* — affirmation of autonomy in the development and governance of technologies, subordinating them to national ethical priorities and culturally grounded conceptions of justice, duty, and progress.

These values form a distinct Russian axiology of responsibility, in which AI ethics is conceived not as a universal set of abstract principles but as a living cultural code that unites technological progress with the self-awareness of community and its values. Such an ethics is capable not only of integrating international standards but also of enriching the global discussion with a new humanist dimension of responsibility in the age of artificial intelligence.

SOVEREIGN AI MODEL

The growing diversification of ethical and political frameworks in global technology governance inevitably raises the question of autonomy — not only moral or cultural, but also technological. If responsibility in a multipolar world is articulated through diverse political ontologies, then artificial intelligence itself becomes a medium through which these differences are institutionalized. In this sense, the emergence of Sovereign AI marks the political and technological dimension of MAR. It reflects the attempt of different societies to align AI development with their own ethical priorities, epistemic traditions, and models of governance. Sovereign AI thus extends the debate on RRI beyond ethics and participation toward the deeper question of who defines, controls, and legitimizes responsibility in a world of competing cultural and geopolitical centers.

The term "Sovereign AI" has gained political and technological significance due to the efforts of Jensen Huang, the CEO of NVIDIA, who publicly articulated this concept between 2023 and 2024 as a response to the threat of global technological dependence. In his interpretation, Sovereign AI refers to the ability of each country to own the production of its own intelligence (Caulfield, 2024). Huang emphasizes that artificial intelligence "codifies your culture, your society's intelligence, your common sense, your history — you own your own data" (ibid.). He insists that the architecture of Sovereign AI must include national AI factories capable of training and deploying large language models controlled by local communities and the state (Lee, 2025). Although the concept of digital and technological sovereignty had been used previously in the media of various countries,

it was Huang who institutionalized and first proposed a comprehensive Sovereign AI architecture as a political-infrastructural project, integrating AI with national identity and autonomy.

Since then, it has been further developed in philosophical-ethical, political, and scientific-technical contexts (Shrier et al., 2024; Dakakni, 2025; Duan et al., 2025; Srivastava & Bullock, 2024). The central aim is the alignment of AI technologies with the values of specific countries and societies. Sovereign AI is not merely a technical priority but a new form of digital self-expression of nations seeking to embed AI into the architecture of their cultural, linguistic, and legal autonomy.

Thus, two key aspects can be distinguished when defining Sovereign AI: a technological and a value-based one. The first pertains to a state's ability to independently develop, maintain, and deploy AI systems with minimal external dependence. This conceptual core is driven by the strategic need to control critical services — from defense to economy — and the desire to avoid dependence on foreign AI platforms and companies. In addition to infrastructural independence, Sovereign AI also implies alignment with national values and ethics: countries aim to define the ethical and social implications of AI themselves, acting in accordance with their cultural and normative specifics, and to prevent the imposition of global standards or external priorities.

In the context of digital transformation, global AI infrastructure is increasingly concentrated in the hands of a limited circle of actors, primarily the USA and China, as well as technological giants controlling cloud platforms, LLM models, computational power, and Big Data. This model is referred to as hegemonic AI: it seeks the universalization of standards, centralization of computations, and the standardization of ethical norms, detached from the political-cultural contexts and interests of individual countries (Carvalho, 2025; Dakakni, 2025). Sovereign AI, on the other hand, emphasizes national control and localization. It asserts that each country has the right to define how the data of its citizens is used, which algorithms are applied, and which values they follow. This manifests in the creation of "sovereign clouds," national data centers, ethical codes, and legal regimes that regulate the use of AI. This model is especially relevant for states aiming to retain autonomy in the face of digital dependence.[3]

[3]Over the past two years, the topic of sovereign AI has become very popular in China. Chinese scholars and media are actively promoting this concept amidst intellectual and geopolitical challenges (Liao & Hong, 2024; Lin, 2025).

The emerging dichotomy reflects a fundamental tension between global integration and digital sovereignty, between the universal rationality of algorithms and the ontological plurality of life-worlds. In other words, it is not only about technological architecture, but the central issue concerns power, legitimacy, and responsibility: who controls the algorithms, who is accountable for their decisions, and in whose interests do these algorithms operate? Unlike hegemonic AI, which imposes norms externally, Sovereign AI embodies a localized political will, thus acquiring legitimacy as a subject of digital power. Sovereign AI, therefore, is not just a set of tools but a form of political-technological organization where algorithmic governance, normative modeling, and sovereign jurisdiction converge. Its implementation requires institutionalized coordination between the state, the scientific community, technology companies, and civil society.

THE CHINESE VIEW OF THE SOVEREIGN AI MODEL: A CRITICAL ANALYSIS

This paper provides a critical analysis of Chinese interpretations of Sovereign AI, as reflected in Chinese chatbot outputs and consolidated within Chinese academic research. Specifically, on May 20, 2025, a report on the concept of Sovereign AI was presented by a research group led by Professor Gu Chao from the School of Government at Beijing University at the "Trustworthy Future: AI Ethics and Societal Transformation" workshop held at the Institute for the History of Natural Science of the Chinese Academy of Sciences. We aim to show how different the interpretations of the Russian model of Sovereign AI are in the context of Russian, Western European, and Chinese sources. In other words, the Russian model of Sovereign AI is presented not only in the West but also in China, often in a one-sided and biased way.

Chinese scholars have developed a typology of Sovereign AI, presenting four main models — American, European, Chinese, and Russian (Table 1).[4] In doing so, they used Chinese chatbots and national concepts for deeper analysis. Even a cursory glance reveals some cultural biases, particularly the excessive "idealization" of the Chinese model and the one-sidedness in describing the Russian model. Let's examine the presented typology in more detail.

USA: MARKET COLLABORATION MODEL

The American model is characterized by a low level of centralized control (the classical "bottom-up" system). AI development in the US is primarily

[4]This table is provided with the kind consent of the author of the report, Prof. Gu Chao (Beijing), who is preparing a corresponding publication based on the presented materials.

driven by private corporations such as OpenAI, Google DeepMind, Meta, and NVIDIA, which set the tone in the global technological race. The government, in turn, performs a strategic regulatory function — through export controls, sanctions, defense agencies (DARPA, DoD), and indirect investments in critical areas. However, the level of network collaboration in the US is very high. Private companies, government institutions, academia, and the venture capitalist sector actively collaborate within a distributed innovation ecosystem. This allows the US to achieve flexibility, speed, and dominance in key technologies, although it also creates problems related to regulatory fragmentation and the lack of a centralized ethical policy.

EUROPEAN UNION: MARKET COLLABORATION MODEL WITH STRONG EMPHASIS ON ETHICS AND REGULATION

The EU model demonstrates a medium level of centralized control. AI governance in Europe is carried out through EU institutions (the European Commission, European Parliament), although each member country retains a certain degree of autonomy. At the core of the model are regulation and human rights: the AI Act, General Data Protection Regulation (GDPR), and other regulations aim to create ethical and transparent AI. Network interaction is also at a high level: not only governments of EU countries and tech companies but also numerous research centers, ethical committees, and civil society representatives are involved in the processes. The EU strives to build an "ethical AI" model, where trust and individual rights take precedence over the speed of innovation.

RUSSIA: THE SOVEREIGNTY-FIRST MODEL

The Russian model of Sovereign AI prioritizes national sovereignty, national security, and strategic autonomy. At its core lies the logic of centralized governance: the state acts not only as a coordinator but also as a key player in AI development — defining the regulatory framework, funding, and implementing AI technologies in critical areas, including defense, intelligence, and cybersecurity. In this model, artificial intelligence is primarily seen as a tool for state control and geopolitical confrontation, rather than as a market- or socially-oriented technology. According to this approach, AI-technologies are mainly geared toward military and defense applications, as well as state governance systems, including facial recognition, public risk prediction, and cyber operations. The private sector plays a subordinate role, often fulfilling state orders or operating within a regulated environment. Market competition and an open innovation ecosystem are underdeveloped, leading to limited commercialization and the introduction of AI in everyday civil

applications. It is important to note that this approach is reinforced by external political factors, primarily sanctions and deglobalization.

CHINA: THE CENTRALIZED COLLABORATION MODEL

The Chinese model of Sovereign AI, referred to as the Centralized Collaboration Model, is a unique blend of strong state leadership and broad participation from various actors — ranging from tech companies to research institutions and regional structures. It is based on a hierarchically organized, networked architecture in which central authority (the Communist Party of China) retains control over key strategic directions, while implementation and development are distributed among partner organizations. This model aims to ensure autonomy in critical technologies, data security, and resilience in the face of external pressures, primarily from the US and its allies. The high level of centralization in this model is expressed through the active role of the state in setting goals, allocating resources, regulatory control, and managing infrastructure. However — and this is a key point — these tasks are implemented through wide, institutionalized collaboration, which involves private corporations, universities, startups, and even local governments.

Country / Region	Degree of Centralized Control	Degree of Network Collaboration	Model Type	Key Features
China	High	High	Centralized Collaboration Model	Government-led, multi-stakeholder collaboration to ensure data security and technological autonomy; gradual expansion of international cooperation
United States	Low	High	Market Collaboration Model	Enterprise-led, market-driven innovation with government policy support; relatively high data security risks
Russia	High	Low	Sovereignty-First Model	Highly centralized government control; focus on military and security technologies; weaker commercialization and application
European Union	Low	High	Market Collaboration Model	Strong emphasis on privacy and ethics; strict data protection frameworks; slower pace of technological innovation

Table 1. Typology of Sovereign AI

In this typology, the Chinese model appears as the most "perfect" and preferred among the four. But is this truly the case? Let's conduct an analysis of the Chinese Sovereign AI model using Western chatbots (ChatGPT) as well as current materials from Russian sources.

Below in Table 2 is the Sovereign AI typology generated by ChatGPT5. The differences between the typologies based on DeepSeek (China) and ChatGPT (USA) are related to varying assessments of the role of network collaboration and the alignment of state policy with international strategy. These differences should be analyzed in more detail:

DEGREE OF CENTRALIZED CONTROL

According to the Chinese typology, the Chinese model is characterized by a high degree of centralized control. This is accurate, as the Chinese government plays an active role in setting strategic directions for technological development and controls key IT companies. The typology generated by ChatGPT also highlights a high degree of centralized control in the Chinese model. Both approaches emphasize the role of the state as a central actor; however, while the American chatbot focuses more on overall state planning and the culturally ideological orientation, the Chinese typology emphasizes technological control and the guiding role of the state in ensuring strategic provision.

DEGREE OF NETWORK COLLABORATION

Here we encounter a key difference in the assessments of the level of network collaboration. The Chinese typology asserts that the Chinese model has a high degree of network collaboration. In this context, Chinese researchers emphasize that within the Centralized Collaboration Model, China actively involves not only government bodies but also private companies, academic institutions, and international partnerships. The focus is on a hybrid model, where the state and the private sector work closely together to ensure technological autonomy, data security, and innovative progress. In the typology based on ChatGPT, however, the Chinese model is characterized by a moderate level of network collaboration, which is attributed to the perception of network collaboration as formally organized and hierarchical. In this second case, the Chinese model is seen as top-down managed; hence, the interaction between the state and private companies is more passive, with private companies playing the role of executors of state will rather than equal partners.

Thus, there is a clear gap between Western and Chinese views on the Chinese Sovereign AI model, which emerges from our comparative analysis. Table 2 exposes active cultural and political biases: Chinese chatbots laud their own Sovereign AI model, whereas the American chatbot offers a more neutral portrayal.

Country / Region	Degree of Centralized Control	Degree of Network Collaboration	Key Features
United States	Low centralization; private sector dominance	High interaction among corporations, academia, and government agencies	Market-coordinated model with strategic control (sanctions, standards, military agencies). Sovereignty realized through infrastructural dominance.
European Union	Medium centralization via EU institutions	Active interaction between Member States, Ethics Councils, and technology companies	Normative-ethical sovereignty based on law, transparency, and protection of citizens' rights. High institutional integration; persistent technical dependence.
China	Very high centralization; state determines goals and architecture	Moderate interaction between state and trusted companies	Centralized state model based on data control, long-term planning, and ideological coordination. AI integrated into social governance.
Russia	Moderate centralization; strong state control	Low networking due to sanctions and institutional weakness	Transitional model aimed at technological autonomy but constrained by limited resources. Attempts to integrate ethics and norms with weak institutionalization.

Table 2. Typology of Sovereign AI: Comparative Analysis of Models of the US, EU, China, and Russia (based on ChatGPT)

CORRECTING THE CHINESE TYPOLOGY OF SOVEREIGN AI: A RUSSIAN PERSPECTIVE

The Chinese Sovereign AI Model: The Centralized System and Party-Corporate Synergy. China represents the most striking example of a highly centralized Sovereign AI model, in which the state, represented by the Communist Party of China (CPC), acts as an algorithmic sovereign in the Hobbesian sense. This is reflected in strategic planning (national AI development

plans, such as the New Generation Artificial Intelligence Development Plan, 2017), centralized control over data (personal information protection law, data security law), and direct government involvement in determining AI research directions. However, despite the strength of centralized control, the Chinese model is not entirely vertical. It actively engages so-called trusted companies (large technology corporations such as Alibaba, Tencent, Baidu, Huawei, iFlytek) that implement the party's strategy in the technological field. These companies act as conduits for state will but also possess high competence, research resources, and international infrastructure.

Thus, the level of network collaboration in China can be characterized as medium (not high level) for the following reasons:

First, the party-corporate synergy, rather than equal partnership. Interaction between the state and private AI actors does not occur based on market contracts, as in the US, or on multi-level political-ethical dialogue, as in the EU, but through party leadership and a mechanism of "joint development," where the state sets the framework, and corporations adapt, striving not for freedom but for stable coexistence with power.

Second, political loyalty as a condition for access to AI development. For example, Baidu's involvement in national projects for developing large language models (Ernie Bot) or Huawei's contribution to cloud infrastructure development is possible only if party lines and strategic guidelines are followed, including censorship, prioritizing the domestic market, and exporting party values to technology.

Third, AI as a tool for social engineering. Key AI developments are embedded in the governance system: from citizen social credit systems to digital surveillance, facial recognition, and biometric control. This implies not just business applications but deep normative integration of AI into the logic of the party-state. In such a system, the private sector functions more as an extension of state will than as an autonomous partner.

Fourth, institutionalized hierarchy in the AI ecosystem. In China, there is no horizontal environment for interaction between academia, the state, and civil society. Instead, there is a clear hierarchy of actors, with ministries and the CPC at the top, while companies are integrated into the overall development plan. This makes the interaction network-like only to some extent.

Thus, the level of network collaboration in China is moderate: it is highly institutionalized but asymmetrical in structure. This is not a Western-style network model but a hierarchical-modular system where coordination and interaction are subordinated to the logic of centralized control. China

demonstrates a kind of technocratic Leviathan, where algorithmic governance permeates state power, and corporations serve as the technological arms of the digital sovereign.

The Russian Sovereign AI Model: A Transitional Model with a Collaborative Orientation. Russia promotes a centrally managed Sovereign AI model, relying on state institutions and corporations. Although in some Chinese analytical reviews, the Russian Sovereign AI model is defined as the Sovereignty-First model — with a high degree of centralization and a low level of network collaboration — this definition does not account for important processes of internal collaboration that have developed in recent years. A particularly telling example in this context is the process of creating and adopting the "AI Ethics Code" in Russia (2021–2024), which became an example of genuinely transdisciplinary dialogue and the involvement of a wide range of stakeholders. The Code itself includes not only declarative provisions but also mechanisms for implementation through the appointment of ethics officers in each signatory organization. This means that the interaction between participants was not limited to the development of the text but became the foundation for a sustainable ethical infrastructure requiring constant communication and joint expertise.

The Code is the result of collaboration among a wide range of stakeholders. Representatives of the government (the Bank of Russia and relevant agencies), the academic community (HSE University, RAS Institute of Philosophy), and the tech industry (Sber, Yandex, and members of the AI Alliance), as well as experts in philosophy, law, and information security, contributed to the document.

> The development of this document became an example of an analytical-deliberative approach, combining scientifically grounded and precautionary strategies; discursive practices were used to assess potential humanitarian consequences and the further development of ethical norms in AI (Maslova et al., 2022: 74).

In other words, the Russian AI model is not reduced to a hierarchical vertical (classical "top-down" system) — instead, it demonstrates signs of institutionalized network collaboration, based on ethical reflection, inter(trans)-disciplinarity, and cooperation between sectors.

If we rely not on abstract indices of digital maturity but on real cases of multilateral participation, Russia's model shows, though not always consistently, the potential to develop collaborative forms of technological sovereignty, where expertise, publicity, and coordination are just as important as centralized management. If we consider real mechanisms of inclusive

regulation, institutional practices, philosophical-normative reflection, and the involvement of multi-actor groups instead of formal external indices, the Russian model demonstrates a significantly higher level of network collaboration than is often assumed in some external classifications. This allows for a rethinking of Russia's trajectory of digital sovereignty as a hybrid form, combining elements of centralized management with institutionalized and academic collaboration.

The vision of MAR provides a useful framework for understanding the coexistence of various models of Sovereign AI. Rather than considering these models as competing or contradictory, the multipolar architecture emphasizes the potential for each to contribute positively to a diversified and balanced global governance system. Each model, with its distinct approach to sovereignty, regulation, and ethical considerations, operates within its own unique political, cultural, and institutional contexts, and collectively, they create a dynamic, multi-layered approach to ethical AI governance.

In this context, MAR is not just a theoretical construct but a call for inclusive collaboration and mutual respect for different regulatory STI-strategies. It envisions a world where countries, rather than imposing a singular global standard, engage in dialogue and adapt their AI policies to local values and needs, while also respecting the shared goals of ensuring security, equity, and sustainability. This vision of a multipolar world is positive in its implications for the development of Sovereign AI, as it encourages a cooperative framework for technological innovation while preventing hegemonic control by any one nation or bloc. In this way, MAR aligns with the principles of RI, emphasizing the need for global ethical norms alongside localized, context-sensitive implementation.

CONCLUSION

This study has argued that the idea of RI, though conceived as a universal framework for aligning technology with ethical and societal values, is deeply rooted in liberal political ideals of participation, deliberation, and transparency. International STS scholarship is searching for a fundamental reorientation of RRI's critical potential, pointing out significant shortcomings of the ethics of technology. The comparative analysis of Chinese and Russian contexts shows that when the RRI policy framework enters non-liberal environments, its ethical grammar transforms. Consequently, the successful implementation of R(R)I in different socio-cultural contexts cannot be a simple "transfer" of a ready-made European model: it is undergoing

a substantive transformation that is highlighted by political-philosophical reflection calling into question normative universalism.

In China, AI ethics is grounded in Confucian ideals of harmony and virtue. Here, moral mediation replaces public deliberation, and social concord functions as the regulative ideal of the good. In Russia, technology is consistently conceived as a part of a moral project oriented toward the collective good, the community's traditions, and technological sovereignty. Together, these trajectories illustrate a different understanding of responsibility than the one accepted in conventional Western RI scholarship: they testify in favor of a multipolar ethics of responsibility, where each civilization strives to articulate its own mode of legitimizing technology and its moral meaning. The concept of MAR offers a model for distinct ethical worlds to cooperate without domination, fostering new forms of solidarity, justice, and humane technological development.

REFERENCES

Blok, V., and P. Lemmens. 2015. "The Emerging Concept of Responsible Innovation: Three Reasons Why It Is Questionable and Calls for a Radical Transformation of the Concept of Innovation." In *Responsible Innovation 2 : Concepts, Approaches, and Applications*, ed. by B.-J. Koops, 19–35. Springer.

Bylieva, D., and A. Nordmann. 2025. "Ontolytic Effects of AI: Widening the Framework for Responsible Research and Innovation." *Philosophy. Journal of Higher School of Economics* 9 (4): 84–104.

Carvalho, V. D. H. de. 2025. "The Race for AI Hegemony." *Socioeconomic Analytics* 3 (1): 1–7.

Caulfield, B. 2024. "NVIDIA CEO: Every Country Needs Sovereign AI Jensen Huang Describes Transformative Potential of Sovereign AI at World Governments Summit." NVIDIA Blog. Accessed Aug. 8, 2025. https://blogs.nvidia.com/blog/world-governments-summit/.

Checketts, L., and B. Chan, eds. 2024. *Social and Ethical Considerations of AI in East Asia and Beyond*. Cham: Springer.

Coeckelbergh, M. 2024. "Artificial Intelligence, the Common Good, and the Democratic Deficit in AI Governance." *AI and Ethics* 5:1491–1497.

Comisso, M. P., B. Gansky, and L. A. Smith. 2024. "Pluralizing RRI Pedagogy: 'Cachando' Tactical Lessons Towards Critical Resistance for Responsible Research and Innovation Learning." *Journal of Responsible Innovation* 11 (1).

Connolly, W. E. 1995. *The Ethos of Pluralization*. Minneapolis: University of Minnesota Press.

Corrêa, N. K., C. Galvão, J. W. Santos, et al. 2023. "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance." *Patterns* 4 (10).

Dakakni, D. 2025. "Artificial Intelligence as a Tool for Data, Economic and Political Hegemony: Releasing the Djinn." *Ethics in Science and Environmental Politics* 25:1–10.

Doezema, T., D. Ludwig, P. Macnaghten, et al. 2019. "Translation, Transduction, and Transformation: Expanding Practices of Responsibility Across Borders." *Journal of Responsible Innovation* 6 (3): 323–331.

Duan, Y., Sh. Huang, and Sh. Gong. 2025. "Strategic Recommendations Report on the Integrated Development of Sovereign AI and Semantic Sovereignty." ResearchGate. Accessed Aug. 8, 2025. https://www.researchgate.net/publication/393464166.

Engel'meyyer, P. K. 1898. *Tekhnicheskiy itog XIX veka [The Technical Outcome of the 19th Century]: opyt filosofii tekhniki [An Essay on the Philosophy of Technology]* [in Russian]. Sankt-Peterburg [Saint Petersburg]: Tipografiya A. S. Suvorina [A. S. Suvorin Printing House].

——— . 2013. *Filosofiya tekhniki [The Philosophy of Technology]* [in Russian]. Sankt-Peterburg [Saint Petersburg]: Lan' [Lan' Publishing House].

"Ethical Norms for New Generation Artificial Intelligence" [in Chinese]. 2021. Ministry of Science and Technology of the People's Republic of China. Accessed Aug. 8, 2025. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

Garbuk, S. V., and A. V. Ugleva. 2024. "Avtomatizirovannyye intellektual'nyye sistemy [Automated Intelligent Systems]: eticheskiy i normativno-tekhnicheskiy podkhody k regulirovaniyu [Ethical and Regulatory Approaches to Regulation]" [in Russian]. *Chelovek [Human]* 35 (4): 98–117.

Godin, B. 2009. "National Innovation System: The System Approach in Historical Perspective." *Science, Technology, and Human Values* 34 (4): 476–501.

González Peralta, P. 2022. "The Ethics of Artificial Intelligence and the Multilateral Push for a Treaty." Universitat Oberta de Catalunya. Accessed Aug. 10, 2025. https://openaccess.uoc.edu/items/b005b374--0fa7--4d15--9cc9-bf1ba3de20c9.

Griessler, E., R. Braun, M. Wicher, and M. Yorulmaz. 2023. "The Drama of Responsible Research and Innovation: The Ups and Downs of a Policy Concept." In *Putting Responsible Research and Innovation into Practice*, ed. by V. Blok. Library of Ethics and Applied Philosophy 40. Cham: Springer.

Grunwald, A., ed. 2024. *Handbook of Technology Assessment*. London: Edward Elgar.

Kamolov, S. G., A. A. Varos, A. Krebits, and M. Yu. Alashkevich. 2022. "Dominanty natsional'nykh strategiy razvitiya iskusstvennogo intellekta v Rossii, Germanii i SShA [Dominants of National Strategies for the Development of Artificial Intelligence in Russia, Germany, and the USA]" [in Russian]. *Voprosy gosudarstvennogo i munitsipal'nogo upravleniya [Public Administration Issues]*, no. 2, 85–105.

"Kodeks etiki v sfere II [Russian AI Ethics Code]" [in Russian]. 2021. Al'yans v sfere iskusstvennogo intellekta [AI Alliance Russia]. Accessed June 1, 2025. https://ethics.a-ai.ru/.

Lee, A. 2025. "What Is Sovereign AI?" NVIDIA Blog. Accessed Aug. 8, 2025. https://blogs.nvidia.com/blog/what-is-sovereign-ai/.

Leushina, V. V., and V. E. Karpov. 2022. "Etika iskusstvennogo intellekta v standartakh i rekomendatsiyakh [Ethics of Artificial Intelligence in Standards and Recommendations]" [in Russian]. *Filosofiya i obshchestvo [Philosophy and Society]* 3 (104): 124–140.

Liao, M., and W. Hong. 2024. "Insights for China from the Sovereign AI Strategies of Other Countries" [in Chinese]. *Economic Prospect*, no. 215, 84–89.

Lin, W. 2025. "Analysis of Sovereign AI Strategies Promoted by Major Countries" [in Chinese]. *Taiwan Economic Research Monthly* 48 (8): 14–22.

Liu, Y., F. Liu, and F. Ma. 2025. "Evaluating Cultural and Linguistic Alignment Across the LLMs." Proceedings of NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle. Accessed Aug. 10, 2025. https://openreview.net/forum?id=S4qF7Yd8ao.

Maslova, Ye. A., N. A. Samoylovskaya, Ye. D. Sorokova, and A. D. Chekov. 2022. "Regulirovaniye iskusstvennogo intellekta v Rossii [Regulation of Artificial Intelligence in Russia]: eklektika podkhodov i aktorov [An Eclecticism of Approaches and Actors]" [in Russian]. *Sravnitel'naya politika [Comparative Politics]* 4 (13): 65–84.

Meier, A., and R. Byland. 2020. *A New Horizon for Society?: Analysing the Integration of Responsible Research and Innovation in Horizon Europe*. Brussels: SwissCore.

Mitcham, C. 2022. "Political Philosophy of Technology: After Leo Strauss (A Question of Sovereignty)." *NanoEthics* 16:331–338.

——— . 2024. "Artificial Intelligence Liberalism: Questions Small and Big." In *Social and Ethical Considerations of AI in East Asia and Beyond*, ed. by L. Checketts and B. Chan, 147–159. Cham: Springer.

Mitcham, C., A. Mikhailovsky, and E. Seredkina. 2024. "Political Philosophy for the Ethics of Artificial Intelligence: A Conversation with American Philosopher Carl Mitcham." *Technologos*, no. 4, 5–18.

Owen, R., P. Macnaghten, and J. Stilgoe. 2012. "Responsible Research and Innovation: From Science in Society to Science for Society, with Society." *Science and Public Policy* 39 (6): 751–760.

Pandey, P. 2024. "Responsible Innovation Goes South: Critique, Othering, and a Commitment to Care." *Journal of Responsible Innovation* 11 (1).

Papadopoulou, D. 2025. "Geopolitics: AI and China: Enabling Ideology?" *Frontiers in Political Science* 7.

Penttilä, L. 2024. "On with Critique! The Necessity of Critique in Addressing the Political Deficits of Responsible Innovation." *Journal of Responsible Innovation* 11 (1).

Qiao-Franco, G., and R. Zhu. 2022. "China's Artificial Intelligence Ethics: Policy Development in an Emergent Community of Practice." *Journal of Contemporary China* 33 (146): 189–205.

Repin, D. A., and S. A. Ignat'yev. 2024. "Vnedryat' nel'zya otkazat'sya [Implementation Impossible to Refuse]: vliyaniye etiki na primeneniye tekhnologiy iskusstvennogo intellekta v upravlenii sotsial'no-ekonomicheskimi protsessami [The Influence of Ethics on Using Artificial Intelligence in Socio-Economic Management]" [in Russian]. *Ekonomika i upravleniye [Economics and Management]* 30 (12): 1503–1509.

Roche, C., P. J. Wall, and D. Lewis. 2023. "Ethics and Diversity in Artificial Intelligence Policies, Strategies and Initiatives." *AI and Ethics* 3:1095–1115.

Rozin, V. M. 2004. "Filosofiya tekhniki [Philosophy of Technology]" [in Russian]. Tsentr gumanitarnykh tekhnologiy [Center for Humanitarian Technologies]. Accessed Aug. 10, 2025. https://gtmarket.ru/library/articles/6309.

Shrier, D. L., A. Piotti, A. Pentland, and A. Faisal. 2024. "Considerations Regarding Sovereign AI and National AI Policy." Trusted AI Alliance and Imperial College London. Accessed Aug. 8, 2025. https://sovereign-ai.org/media/papers/Executive_Summery_Sovereign_AI.pdf.

Srinivas, K. R. 2022. "Responsible Research and Innovation and India: A Case for Contextualization and Mutual Learning." In *Ethics, Integrity and Policymaking : The Value of the Case Study*, ed. by D. O'Mathúna and R. Iphofen, 29–48. Research Ethics Forum 9. Cham: Springer.

Srivastava, S., and J. Bullock. 2024. "AI, Global Governance, and Digital Sovereignty." arXiv. Accessed Aug. 10, 2025. https://arxiv.org/abs/2410.17481.

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580.

"Ukaz Prezidenta Rossiyskoy Federatsii ot 26.10.2020 № 645 'O Strategii razvitiya Arkticheskoy zony Rossiyskoy Federatsii i obespecheniya natsional'noy bezopasnosti na period do 2035 goda' [Decree of the President of the Russian Federation No. 645 of October 26, 2020, 'On the Strategy for the Development of the Arctic Zone of the Russian Federation and Ensuring National Security through 2035']" [in Russian]. 2020. Ofitsial'noye opublikovaniye pravovykh aktov [Official Publication of Legal Acts]. Accessed June 1, 2025. http://publication.pravo.gov.ru/Document/View/0001202010260033.

Ulnicane, I. 2021. "Artificial Intelligence in the European Union: Policy, Ethics and Regulation." In *Routledge Handbook of European Integrations*, ed. by Th. Hoerber, G. Weber, and I. Cabras, 254–269. London: Routledge.

UNESCO. 2021. "Ethics of Artificial Intelligence: The Recommendation." UNESCO. Accessed June 1, 2025. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

Vasil'yev, S. N. 2022. "Iskusstvennyy chelovek i obshchestvo [Artificial Human Being and Society]" [in Russian]. In *Chelovek i sistemy iskusstvennogo intellekta [Human Being and Artificial Intelligence Systems]*, ed. by V. A. Lektorskiy, 30–58. Sankt-Peterburg [Saint Petersburg]: Aleteyya.

Von Schomberg, R. 2013. "A Vision of Responsible Research and Innovation." In *Responsible Innovation : Managing the Responsible Emergence of Science and Innovation in Society*, ed. by R. Owen, M. Heintz, and J. Bessant, 51–74. London: Wiley.

——— . 2015. "Responsible Innovation." In *Ethics, Science, Technology, and Engineering : A Global Resource*, ed. by J. B. Holbrook and C. Mitcham. Farmington Hills (MI): Gale / Cengage Learning.

——— . 2025. "On Technological and Innovation Sovereignty: A Response to Carl Mitcham's Call for a Political Theory of Technology." *Nanoethics* 19 (2).

Wang, L., and T. B. Long. 2023. "The Conceptual Evolution of Responsible Research and Innovation in China: A Systematic Literature Review." *Journal of Responsible Innovation* 10 (1).

Wong, P. H. 2016. "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" *Journal of Responsible Innovation* 3 (2): 154–168.

——— . 2024. "Confucian 'Trustworthy AI': Diversifying a Keyword in the Ethics of AI and Governance." In *Social and Ethical Considerations of AI in East Asia and Beyond*, ed. by L. Checketts and B. Chan, 3–14. Cham: Springer.

Wong, P. H., and T. X. Wang, eds. 2021. *Harmonious Technology: A Confucian Ethics of Technology*. Routledge.

Wu, P., G. Shen, D. Zhao, et al. 2025. "CVC: A Large-Scale Chinese Value Rule Corpus for Value Alignment of Large Language Models." arXiv. Accessed Aug. 10, 2025. https://arxiv.org/abs/2506.01495v4.

Zhao, Y., and M. Liao. 2019. "Chinese Perspectives on Responsible Innovation." In *International Handbook on Responsible Innovation. A Global Resource*, ed. by R. Von Schomberg and J. Hankins, 426–440. Cheltenham: Edward Elgar.

Mikhailovsky A. V., Seredkina E. V. [Михайловский А. В., Середкина Е. В.] Political Philosophy of Technology and Responsible Innovation in a Multipolar World [Политическая философия техники и ответственные инновации в многополярном мире] : The Russian and Chinese Cases of AI Ethics [российский и китайский подходы к этике искусственного интеллекта] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 13–46.

Александр Михайловский
д. филос. н., доцент, Национальный исследовательский университет «Высшая школа экономики» (Москва); orcid: 0000–0001–9687–114x

Елена Середкина
к. филос. н., доцент, Пермский национальный исследовательский политехнический университет (Пермь); orcid: 0000–0003–2506–2374

Политическая философия техники и ответственные инновации в многополярном мире

российский и китайский подходы к этике искусственного интеллекта

Получено: 10.10.2025. **Рецензировано:** 11.11.2025. **Принято:** 16.11.2025.

Аннотация: Данная статья предлагает политико-философский подход к концепции ответственных инноваций (ОИ) и этического управления искусственным интеллектом (ИИ) в многополярном мире. В статье утверждается, что хотя ОИ обычно понимается как нейтральная и универсальная модель для согласования науки и техники с этическими и общественными ценностями, эта дискурсивная рамка остается глубоко укорененной в западных либерально-демократических представлениях о делиберативной политике, инклюзии и прозрачности. При попытке применить эти принципы в нелиберальных культурных и политических традициях они претерпевают существенные трансформации, указывая на ограничения нормативного универсализма. На основе сравнительного анализа китайского и российского подходов статья ставит под вопрос западные либерально-демократические основы ОИ и этического управления ИИ. Китайская модель, уходящая корнями в конфуцианские категории гармонии и добродетели, трактует ответственность как «моральную культуру», в то время как российский подход усматривает связь ОИ с общим благом и технологическим суверенитетом. В работе критикуется асимметричное восприятие обеих моделей управления ИИ, где китайская видится как централизованная, но гармоничная, а российская — как централизованная и государство-центричная, и предлагается новое видение российского «суверенного ИИ» как системы сотрудничества, обеспечивающей диалог между государством, бизнесом и наукой: российский Кодекс этики ИИ (2021) является ярким примером междисциплинарного и межсекторального взаимодействия. Наконец, в статье вводится понятие многополярной архитектуры ответственности (МАО), которое обозначает пространство для реализации культурного разнообразия наций внутри общей гуманистической перспективы. Эта рамочная конструкция представляет науку и технологии как инструменты глобального сотрудничества, а не геополитической конкуренции.

Ключевые слова: политическая философия техники, ответственные инновации (ОИ), ответственные исследования и инновации (ОИИ), многополярный мир, многополярная архитектура ответственности (МАО), искусственный интеллект (ИИ), этика искусственного интеллекта, суверенный искусственный интеллект.

DOI: 10.17323/2587-8719-2025-4-13-46.

Wang, D. 2025. "From Engineering Ethics to Ethical Engineering: Leveraging AI for Governing Emerging Technologies" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 47–67.

Dazhou Wang*

From Engineering Ethics to Ethical Engineering**

Leveraging AI for Governing Emerging Technologies

Submitted: Sept. 16, 2025. **Reviewed:** Nov. 11, 2025. **Accepted:** Nov. 11, 2025.

Abstract: Ethical Engineering (EtEn) is an emerging discipline that represents a paradigm shift from traditional Engineering Ethics (EnEt). Rather than focusing primarily on educating individual practitioners, EtEn aims to systematically embed ethical principles into the very fabric of technological systems and governance processes. This paper examines this fundamental transition from EnEt, which focuses on educating practitioners about normative principles, to EtEn, which treats ethics as a systematic engineering problem, focusing on translation of principles into executable governance tools. The study highlights AI's dual role as both the primary domain requiring governance and a pivotal enabler for it, examining its potential to enhance ethical governance through improved algorithmic auditability, support for complex ethical decision-making, and cross-domain collaborative governance, while also addressing challenges like value alignment, bias mitigation, and technological reductionism. It identifies eight key issues that constitute the core research agenda for EtEn and argues that its development must be understood as an experimental, iterative process. This paradigm shift not only expands practical pathways for implementing EnEt but also offers novel methodological support for the ethical governance of emerging technology in the age of AI.

Keywords: Engineering Ethics, Ethical Engineering, Artificial Intelligence (AI), Emerging Technology, Value Sensitive Design, Moralizing Technology.

DOI: 10.17323/2587-8719-2025-4-47-67.

1. INTRODUCTION

Technological innovation in artificial intelligence (AI), biotechnology, nanotechnology, and neurotechnology is advancing rapidly, bringing unprecedented opportunities and formidable societal challenges. These technologies are complex, uncertain, and transformative, often stretching traditional ethical and governance mechanisms to their limits. Existing oversight models

*Dazhou Wang, PhD in Philosophy; Professor at the School of Humanities, University of Chinese Academy of Sciences (UCAS) (Beijing, China), dzwang@ucas.ac.cn, ORCID: 0000–0001–9586–4597.

Acknowledgements: This research was funded through the Philosophy and Social Sciences Major Program of Ministry of Education of China (23JZD006).

tend to be reactive and slow, struggling to anticipate, evaluate, and mitigate ethical risks in a timely manner.

For decades, the primary response has been engineering ethics (EnEt), which emphasizes educating engineers and technologists in ethical principles (Harris Jr. et al., 2013; Martin & Schinzinger, 2010). This approach relies on professional codes of conduct, case studies, and individual moral reasoning. While important, it operates mainly at the individual or organizational level, depending heavily on human judgment and self-regulation, which can be subjective, variable, and difficult to scale across global and distributed technology ecosystems. It often functions as an external constraint rather than an integrated component of the engineering process.

Currently, a new paradigm is emerging: Ethical Engineering (EtEn). While several books use the term "Ethical Engineering" (Hersh, ed., 2015; Schlossberger, 2023), they primarily operate within the established paradigm of EnEt, focusing on helping engineers understand and address ethical issues. In contrast, this paper articulates and defends a distinct conception of EtEn as a novel engineering science (Wang, 2023). Its core thesis is that EtEn represents a paradigm shift from advising individual engineers to systematically building ethics into engineering systems and processes, with AI serving as a key enabler for making responsible engineering achievable at scale. This direction is reflected in research exploring how to embed ethical considerations into autonomous systems (Wallach & Allen, 2009), develop tools for ethical risk reflection (Urquhart & Craigon, 2021), and extend value sensitive design (Friedman & Hendry, 2019; Friedman et al., 2013) across the AI lifecycle (Umbrello & van de Poel, 2021). While EnEt asks, "How can engineers behave ethically?", EtEn asks, "How can we design systems, processes, and tools that actively enable ethical outcomes?" This reframing shifts the goal from using moral philosophy to constrain practice to systematically embedding ethical reasoning into the fabric of technological systems and governance structures.

At the heart of this emerging paradigm lies AI — a domain fraught with ethical questions yet also a potent source of solutions to complex governance problems. While many scholars and practitioners focus on the ethical governance of AI development, a compelling strand of research emphasizes AI's capacity to function as an instrument of governance. This is reflected in diverse applications such as managing dual-use technologies (Ulnicane et al., 2023), countering cyber threats (Camacho, 2024), improving regulatory adherence (Jain et al., 2024; Padmanaban, 2024), and combating corrupt practices (Adobor & Yawson, 2023). A growing consensus suggests that AI

can provide an integrative framework for navigating the intricate challenges at the intersection of technological innovation and societal evolution.

This paper aims to systematically develop the conceptual foundations of Ethical Engineering as a distinct paradigm. First, it outlines the conceptual transition from EnEt to EtEn, including a clarification of EtEn's unique position relative to value sensitive design (VSD) and the moralizing technology (MT) approach (Verbeek, 2006; 2011). Second, it examines AI's dual role in EtEn — as both its primary testbed and most powerful accelerator — and analyzes the strategy of "using AI to govern AI." Third, it identifies eight key challenges that define the research agenda for EtEn. Finally, it concludes by emphasizing the need for a reflexive, experimentalist approach to developing EtEn as a socio-technical governance system, one that requires interdisciplinary collaboration and cross-cultural sensitivity.

2. FROM ENGINEERING ETHICS TO ETHICAL ENGINEERING: A PARADIGM SHIFT

The field of technology ethics is undergoing a significant evolution, marked by a transition from the established framework of EnEt to the emerging paradigm of EtEn. This shift, facilitated by intermediary approaches like VSD and the theory of MT, represents a move from principle-based guidance toward system-based implementation.

2.1. CHARACTERISTICS OF ENGINEERING ETHICS

EnEt has long been an established field dedicated to addressing the moral obligations and dilemmas that engineers encounter in their practices. At its core, it is built upon several key tenets. Normative principles form the foundation of EnEt. Organizations such as the National Society of Professional Engineers (NSPE) and the Institute of Electrical and Electronics Engineers (IEEE) have developed codes of ethics that emphasize public health, safety, and welfare. Philosophical frameworks like utilitarianism and deontology also play a crucial role in guiding ethical reasoning and decision-making. While utilitarianism focuses on maximizing overall well-being, deontology emphasizes adherence to moral rules. More recently, the importance of virtue ethics is gaining increasing recognition in the field of engineering ethics.

Education and individual agency are also central to EnEt. It places strong emphasis on teaching students and professionals how to recognize ethical issues, analyze complex dilemmas, and make morally sound decisions. Through

courses and training programs, engineers are equipped with the necessary skills to cope with the ethical challenges they may face in their careers.

Another characteristic of traditional EnEt is its case-based and reactive nature. Historical cases, such as the Challenger disaster, are often used as teaching tools (Elliot et al., 1993). These cases provide valuable lessons about the consequences of unethical practices. However, this approach is retrospective, as it primarily analyzes past events in order to prevent similar mistakes in the future.

EnEt also functions as an advisory and constraining framework. It sets out a set of rules and guidelines that engineers must follow to ensure that their work remains within socially acceptable boundaries. It acts as an external force that guides engineering practice, but it does not necessarily provide a comprehensive solution to all modern engineering challenges.

Despite its importance, the traditional paradigm of EnEt faces several modern challenges. One of the main issues is limited scalability. As technology develops at an ever-increasing pace, it becomes difficult to apply a one-size-fits-all set of ethical principles to a wide range of engineering projects. Weak enforcement is another problem. There is often a lack of strict mechanisms to ensure that engineers adhere to ethical guidelines. Moreover, the fast-paced nature of agile software and technology development is not well-matched with the relatively slow and static nature of traditional EnEt.

2.2. CHARACTERISTICS OF ETHICAL ENGINEERING

In contrast, EtEn is an emerging paradigm that builds on EnEt but differs from it in fundamental ways. It constitutes a distinct engineering discipline focused on designing systems that inherently facilitate ethical outcomes. EtEn is proactive and procedural. It seeks to integrate ethics into every stage of the development lifecycle, from research and development to deployment and decommissioning. By incorporating tools, processes, and checkpoints at each stage, it ensures that ethical considerations are not an afterthought but an integral part of the engineering process.

This new paradigm is also system-oriented. It considers the entire sociotechnical system, including the interactions among humans, technology, and institutions. Instead of focusing solely on the actions of individual engineers, it looks at how the entire system functions and how ethical issues can arise from these complex interactions.

As is well known, a core challenge in EtEn is the technical translation of values. Abstract ethical principles such as fairness, transparency, and

accountability need to be translated into concrete, measurable technical requirements (Wang, 2020). For example, ensuring fairness in an algorithm may require developing specific metrics and techniques to detect and correct biases. EtEn is enabling and empowering. Rather than merely constraining the behavior of engineers, it provides tools and resources that empower developers and governance bodies. Bias detection systems can help identify and mitigate biases in algorithms, while ethical checklists can guide engineers through the ethical decision-making process. These tools allow engineers to build more trustworthy products and services.

IBM has provided a specific example of EtEn. AI Fairness 360, an open-source toolkit developed by IBM, provides a standardized "toolbox" containing dozens of algorithmic fairness metrics (such as demographic parity and equalized odds) and bias mitigation algorithms (such as reweighting and adversarial debiasing). Engineers can seamlessly integrate AIF360 into their machine learning workflows by simply importing it like any other library. They can then use different metrics to calculate their model's fairness scores across various demographic groups (such as different genders or races), and experiment with different debiasing algorithms to determine which method most effectively enhances fairness while maintaining model accuracy.

This shift from EnEt to EtEn can be compared to the introduction of quality assurance (QA) in manufacturing (Feigenbaum, 2012). Just as QA introduced systematic processes to ensure that quality was built into products, EtEn aims to systematically integrate ethics into engineering system. It represents a new way of thinking about engineering, one that is more proactive, comprehensive, and better suited to the challenges of the twenty-first century.

2.3. COMPARISON BETWEEN ENGINEERING ETHICS AND ETHICAL ENGINEERING

The evolution from EnEt to EtEn signifies a shift from principle-based guidance for individuals to the systematic, engineering-based implementation within systems (Table 1). This transformation turns ethical considerations from an external constraint into an intrinsic element of technological design and development. This transition represents a necessary paradigm shift for addressing the limitations of traditional EnEt in the face of modern engineering's complexity and pace. EtEn offers a more holistic and proactive solution, enabling engineers to build technologies that are not only innovative but also ethical and trustworthy. As such, it defines a new frontier for the engineering sciences.

Aspect	Engineering Ethics	Ethical Engineering
Core Focus	Moral obligations and decision-making of individual engineers	Building technological systems and processes that facilitate ethical outcomes
Methodology	Principle-oriented: Based on normative ethical principles (e. g., utilitarianism, deontology) and professional codes of conduct.	System-oriented: Treats ethics as an engineering problem, implemented through designed processes, tools, and system specifications.
Primary Goal	Education and practice constraints: Cultivate engineers' moral reasoning abilities and constrain their behavior within socially acceptable boundaries.	Systematic implementation and empowerment: "Hardwire" ethics into the entire development lifecycle, providing developers with tools and methods to achieve ethical goals.
Temporal Orientation	Post-hoc reflection and reactive: Primarily involves teaching and reflection through the analysis of historical cases.	Proactive and forward-looking: Integrate ethical considerations from the initial R&D phase through deployment and decommissioning.
Core Challenge	Addressing ethical dilemmas faced by individuals, emphasizing personal professional responsibility.	The "translation problem": Converting abstract ethical principles (e. g., fairness, transparency) into concrete, measurable, and implementable technical requirements and system specifications.
Mechanism of Action	Acts as an external constraint: Functions as rules and guidelines that constrain engineering practice.	Acts as an internal enabler: Creates tools and processes (e. g., ethics checklists, bias detection tools) to actively empower developers and governance bodies.
Level of Concern	Individual level: Focuses on the agency and responsibility of the engineer.	System level: Concerns the entire socio-technical system of interactions between people, technology, and institutions.

Table 1. Engineering Ethics vs. Ethical Engineering

2.4. VALUE SENSITIVE DESIGN AND MORALIZING TECHNOLOGY AS BRIDGING CONCEPTS

The transition from EnEt to EtEn is logically and methodologically facilitated by two key frameworks: MT and VSD. They serve as indispensable conceptual and methodological bridges, respectively.

EnEt, while essential, often operates reactively, focusing on educating individual engineers to reason about dilemmas using normative principles and historical cases. However, it struggles to provide scalable, procedural methodologies for integrating ethics into modern technological development, which is increasingly fast-paced and complex. This gap is precisely where VSD and MT intervene, paving the way for EtEn.

Logically speaking, the theory of MT provides the essential philosophical bridge. It challenges the view of technology as a neutral tool by arguing that artifacts actively shape moral decisions and behaviors, suggesting they can even exhibit a form of "moral agency." This perspective fundamentally expands the scope of ethical concern from the individual engineer to the technology itself and its socio-technical context. It provides the fundamental "why" for EtEn: because technology shapes morality, we must consciously design that influence.

Concurrently, VSD provides the crucial methodological bridge. It moves beyond abstract deliberation by offering a structured, tripartite methodology (conceptual, empirical, and technical investigations) for proactively identifying and embedding human values into technical design. This process-oriented framework is a direct precursor to EtEn, which seeks to systematize and operationalize such processes across the entire development lifecycle. VSD translates values such as privacy and fairness into tangible design requirements, directly addressing the "translation problem" that lies at the center of EtEn.

Thus, we can conceptualize their relationship as a continuum from philosophical foundation to technical execution: MT (Philosophical Layer) → Value Sensitive Design (Methodological Layer) → EtEn (Engineering & Implementation Layer) (Table 2). MT justifies the need for EtEn; VSD provides a key methodological process for it; and EtEn is the engineering discipline that systematizes, executes and governs the integration of their insights. For example, VSD, through stakeholder analysis, can determine what "fairness" should mean in a specific context, and EtEn is responsible for technically implementing this defined "fairness" through algorithms and system architectures.

DIMENSION	MORALIZING TECHNOLOGY	VALUE SENSITIVE DESIGN	ETHICAL ENGINEERING
Theoretical Level	Philosophical Foundation	Design Methodology	Engineering discipline & Technical Implementation
Core Focus	Revealing the moral relationship between technology and humans, explaining how technological artifacts mediate and shape human moral perception, decision-making, and behavior.	Providing a systematic design process to proactively incorporate human values into technology design and development.	Establishing a systematic engineering discipline for translating ethical principles into concrete, computable rules and algorithms embedded in technological systems and processes.
Central Question	Do technological artifacts themselves have moral significance? How do they actively influence morality?	How should we systematically consider and embed human values in design?	How should we systematically build and govern technical systems to ensure they produce ethical outcomes?
Theoretical Aim	Descriptive & Explanatory: Describing and explaining the phenomena and mechanisms of technological mediation effects.	Prescriptive: Providing a guiding framework and tools for "what should be done."	Constructive & systematic: Providing the principles, tools, and standards for "how to systematically build and govern" ethical technology.
Understanding of "Ethics"	Ethics is a relational product emerging from human-technology interaction. Morality is "materialized" in the materiality of technology.	Ethics refers to human-centric values that need to be identified and coordinated (e. g., privacy, fairness, well-being).	Ethics is a system property achievable through engineering practices, involving formalizable, operationalizable rules and constraints.
Primary Methods	Philosophical speculation, case studies (e. g., speed bumps, ultrasound machines).	Tripartite iterative investigations (conceptual, empirical, technical); stakeholder analysis.	Formal methods, algorithm design, verification and validation, safety engineering, standard setting.
Focus on Agency	Emphasizes the "moral agency" of technology itself (non-human intentional influence achieved through function).	Emphasizes the agency of humans (designers, stakeholders), who actively make value choices and embeddings.	Emphasizes the agency of the engineering system and process to reliably achieve ethical outcomes.

continued on next page

continued from previous page

Dimension	Moralizing Technology	Value Sensitive Design	Ethical Engineering
Temporal Orientation	Reflective & Prospective: Analyzes existing technologies, and its insights also guide future design.	Proactive: Primarily applied at the beginning and throughout the technology design process.	Proactive & Concomitant: Implemented during the design phase and continuously executed during system operation.
Typical Applications	Explaining how social media mediates interpersonal relationships; explaining how autonomous vehicles alter concepts of responsibility.	Designing privacy-respecting web browser cookie notices; designing urban planning software that supports democratic deliberation.	Designing the full stack of governance for an AI system, from ethical checklists and bias detection in development to real monitoring and audit trails in deployment.

Table 2. Comparison between Value Sensitive Design, Moralizing Technology, and Ethical Engineering

In summary, MT and VSD are not superseded by EtEn but are foundational to it. A robust approach to the ethical governance of technology requires their integration: beginning with the perspective of MT to recognize the moral influence of technology; proceeding to the methodology of VSD to organize and guide the design process; and culminating in the systematic engineering discipline of EtEn to achieve the defined ethical goals reliably and at scale.

3. Artificial intelligence: the "accelerator" and "proving ground" of ethical engineering

AI, particularly Machine Learning (ML) and Deep Learning (DL), possesses unique technical characteristics that make it not only the most critical application domain for EtEn but also a core catalyst and enabling tool that drives the maturation of its methodologies and sharpens the precision of its theories.

3.1. Artificial intelligence as the "accelerator" of ethical engineering

As an enabling tool, AI technology can significantly enhance the effectiveness, precision, and scale of the EtEn toolbox, propelling it from manual, qualitative analysis into a new stage of automation, quantification, and

systematization. AI-driven static and dynamic analysis tools can automatically scan codebases and training datasets to identify potential patterns of bias, security vulnerabilities, or privacy-leakage risks, thereby overcoming the limitations of traditional manual code reviews when dealing with complex systems comprising millions of lines of code. For instance, tools can automatically detect representational biases in datasets or evaluate a model's performance disparities across different demographic subgroups, providing engineers with quantified fairness reports.

Furthermore, EtEn emphasizes the prospective assessment of technological consequences. AI-Enhanced Agent-Based Simulation can construct highly complex virtual social environments to deploy and test algorithms or systems, observing the emergent behaviors and long-term ripple effects generated through interactions with vast numbers of simulated users. This allows engineers to "rehearse" potential unintended ethical consequences of a system, such as the formation of information cocoons, market manipulation, or the exacerbation of social discrimination, at a lower cost before real-world deployment.

Realizing "value alignment" also presents a major challenge: how to extract actionable design inputs from diverse and sometimes conflicting human preferences (Gabriel, 2020). AI techniques, particularly Inverse Reinforcement Learning and advanced interview analysis, can help systematically learn and infer underlying value preferences from human behavior, decisions, or feedback, formalizing them into reward functions or constraints, thereby offering a data-driven engineering path for value alignment. Explainable AI (XAI) is not merely a goal for enhancing model transparency; it is itself a crucial tool for implementing EtEn (Adadi & Berrada, 2018; Dwivedi et al., 2023). Only when a system's decision-making process can be explained and traced can engineers and auditors effectively diagnose it for unfairness, discrimination, or logical errors. For instance, SHAP (SHapley Additive exPlanations) is a leading technique in XAI that explains the output of any machine learning model by quantifying the contribution of each input feature to a single prediction. As a proto-ethical-engineering tool, SHAP has the potential to operationalize the ethical principle of transparency, provide the foundational capability for auditability, and ultimately transforms an opaque "black box" model into a system that can be interrogated and understood. Thus, developing and integrating XAI tools is a core part of building trustworthy, auditable AI systems and an indispensable component of EtEn methodology.

Additionally, ethical norms and social standards are constantly evolving. AI systems can be used for the continuous monitoring of deployed products, analyzing user feedback, public discourse, and operational data in real-time to automatically detect whether their behavior is beginning to deviate from established ethical guidelines or newly enacted laws and regulations, enabling dynamic compliance management and early warning.

3.2. ARTIFICIAL INTELLIGENCE AS THE "PROVING GROUND" OF ETHICAL ENGINEERING

As the core object, the complexity, uncertainty, and autonomy of AI technology pose unprecedented challenges for EtEn, which in turn powerfully drives the discipline's deepening and maturation. It forces the mathematization and operationalization of ethical principles.

Traditional EnEt often deals with principled but vague concepts. However, confronted with AI algorithms, we must provide mathematical definitions for concepts like "fairness": Is it equality of opportunity or equality of predictive outcomes? This pressing engineering necessity forces philosophers, legal scholars, and social scientists to collaborate with engineers in translating abstract ethical concepts into computable, optimizable, and trade-off-able engineering metrics. Without the challenge posed by AI, the refined discussions of "fairness" and "accountability" within EtEn would not have reached their current depth.

AI also highlights the importance of systematic ethics. AI systems are typically not isolated models but components embedded within vast sociotechnical systems. Their ethical impact is often not determined by a single algorithm, but is an emergent property arising from the interaction of multiple stages: data collection, feature engineering, model training, deployment environment, and user interaction. This forces EtEn to develop a system-level perspective and methodology, requiring ethical assessment and governance across the entire system lifecycle rather than focusing solely on the "materialization" during the design phase.

Moreover, application of AI has spawned an urgent global need for AI governance, directly promoting progress in EtEn regarding standard development, certification processes, and the creation of audit tool. AI acts as a "stress test," examining and accelerating EtEn's evolution from corporate self-regulation toward industry regulation and societal governance. As a good example, IEEE has launched the renowned IEEE 7000™ series of standards, particularly "IEEE 7000–2021: Standard Model Process for Addressing Ethical Concerns during Systems Design." It provides a concrete

methodology for directly integrating ethical considerations into systems engineering processes, requiring the identification and managment of ethical risks early in the design phase.

AI and EtEn are in a profound symbiotic relationship. On one hand, AI is the most severe challenge and the most important "proving ground" for EtEn. With its extreme complexity and social impact, it exposes the shortcomings of traditional ethical thinking and urgently demands a rigorous engineering solution, thereby powerfully driving the emergence and development of EtEn. On the other hand, AI technology is also the most powerful "enabler" for EtEn. It provides powerful tools such as automated analysis, large-scale simulation, and preference learning, making the systematic implementation of ethical design, assessment, and monitoring possible and thereby promoting the implementation and evolution of EtEn methodologies.

Therefore, AI is both the primary object of governance and the core means for achieving the governance objectives. Such dual nature makes the field of AI the most active, cutting-edge, and methodology-intensive area for EtEn thought and practice. Advancing AI ethics is, in essence, the practice of building and developing EtEn itself. The two complement each other, working together toward the core goal of ensuring that those increasingly autonomous, powerful, and ubiquitous technologies can robustly, reliably, and responsibly serve human well-being.

3.3. ENHANCING ETHICAL ENGINEERING BY USING AI TO GOVERN AI

The accelerating sophistication and integration of AI across social and economic systems has ushered in an era of unprecedented potential, yet it also introduces profound ethical and governance challenges. Traditional oversight mechanisms, often reliant on slow and subjective human review, are increasingly inadequate for regulating autonomous, large-scale, and rapidly evolving AI systems. In response, a promising yet complex paradigm is drawing attention: the idea of using AI itself to govern AI. This approach represents a fundamental shift within EtEn, moving away from external, intermittent checks toward embedded, continuous, and automated oversight. By leveraging AI's capabilities to monitor, evaluate, and even correct other AI systems, we introduce a layer of reflexivity, a capacity for self-awareness and adaptation, that is essential for managing the ethical risks of advanced technologies (Gou et al., 2023; Madaan et al., 2023; Collin et al., 2023).

The appeal of AI-driven governance lies in its ability to operate at the speed and scale of the systems it monitors. Unlike human committees, AI supervisors can analyze millions of decisions in real time, detect subtle

patterns of bias or malfunction, and respond instantaneously to deviations. This enables a shift from post-hoc auditing to proactive ethical assurance. Core to this approach are several technical pathways that embody this reflexivity. One is Ethics-by-Design, which involves formalizing ethical principles into computational metrics, such as fairness definitions or privacy constraints, that can be built directly into AI architectures. Another is real-time monitoring through multi-agent systems, where guardian AI agents observe the behavior of primary models, flagging anomalies such as discriminatory outputs or performance decay. Furthermore, these systems can enable dynamic self-correction, allowing AI to adjust its own operations in response to ethical breaches. Simulation tools add another dimension, permitting the forecasting of long-term societal impacts before deployment, while blockchain-based audit trails create immutable records for accountability.

However, using AI to govern AI is not a straightforward solution. It introduces a series of deep and potentially recursive ethical and technical complications. The most fundamental is the meta-ethical dilemma: Who decides which values embedded in the governance AI? Ethical norms vary across cultures and jurisdictions, and encoding a single universal standard risks cultural imposition or ethical simplification, a challenge that echoes the value operationalization difficulties highlighted in the previous section. Moreover, AI systems today lack the nuanced understanding required to interpret context-rich moral dilemmas; their strength lies in quantifying metrics, not in interpreting philosophical nuance. This technical limitation becomes especially salient in edge cases, where rigid rules may fail. Additionally, governance systems are themselves vulnerable to adversarial attacks because malicious actors may manipulate supervision mechanisms, bypass safeguards, or poison the training data of the guardian AI. Beyond these risks, there are practical barriers related to standardization and cost. Without interoperable frameworks and shared standards, AI governance may remain fragmented across regions and industries. Meanwhile, the high expense of developing advanced oversight tools could exclude smaller entities, widening the gap between ethical haves and have-nots.

Looking ahead, the future of reflexive AI governance will depend on coordinated efforts across multiple domains. Critical to this effort is the development of open-source tools and benchmarks that make ethical oversight more accessible and reproducible. Equally important is sustained interdisciplinary collaboration through which ethicists, social scientists, engineers, and policymakers must work together to ensure that governance systems are

both technically robust and socially legitimate. An incremental implementation pathway is advisable, one that retains meaningful human oversight, particularly for high-stakes decisions, while gradually introducing greater automation as the technology proves reliable. Finally, international coordination will be essential to avoid regulatory fragmentation and to foster alignment on core ethical principles, even as technical approaches may vary.

4. KEY ISSUES THAT ETHICAL ENGINEERING MUST FOCUS ON

The preceding discussion of using AI to govern AI underscores a fundamental characteristic of EtEn: its inherently experimental nature. The idea that EtEn could provide definitive, universal solutions is a misconception. It is better understood as a continuous process of experimentation between technology and ethics (Wang, 2018; Van de Poel, 2020). Through iterative learning and adaptation, this process gradually aligns technological development with ethical principles and human well-being. However, as an emerging field, its path of development is far from smooth; a series of profound and complex core issues urgently require exploration and resolution. The extent to which these problems are solved will directly determine whether EtEn can transition from a theoretical concept to a mature practice, truly fulfilling its mission of shaping responsible technology. These challenges constitute the core research agenda for EtEn as a discipline.

First, the Problem of Value Operationalization and Quantification. Ethical values such as "fairness," "privacy," "autonomy," and "security" are inherently abstract, qualitative, and highly context-dependent. The primary task of EtEn is to provide engineers with a set of methods to transform these "soft" values into "hard" technical parameters that can be understood, encoded, measured, tested, and optimized. This translation poses significant challenges, as illustrated by the concept of "fairness": the field of algorithms offers dozens of mathematical definitions, such as statistical parity, equality of opportunity, and individual fairness, each carrying distinct philosophical assumptions and legal implications. The critical question then becomes which definition should be selected for a given context, and on what normative grounds? Similarly, operationalizing "privacy" requires designing computable metrics, akin to a "loss function," that can be balanced against other optimization goals like accuracy and latency. These challenges are further compounded when deploying technologies globally, where divergent cultural norms and legal frameworks demand adaptable implementations of core values.

Second, the Problem of Ethical Emergence in Complex Systems. The ethical issues of modern technological systems (e. g., smart cities, platform ecosystems, the Internet of Things) are often emergent properties arising from the interactions among system components, rather than simple summations of the attributes of individual algorithms or modules. This leads to significant difficulties in prediction and governance (Brey, 2012). A typical problem is "compound unfairness": a fair recommendation algorithm combined with a fair pricing algorithm may still produce systematic price discrimination or service exclusion against a particular group across the entire ecosystem. Such long-tail, cross-domain chain reactions cannot be fully anticipated at the design stage. Therefore, EtEn must move beyond the moralization of single technologies and develop theories and tools for system-level ethical simulation, real-time monitoring, and intervention. How to build digital twin environments that can simulate complex human-computer interactions to predict the ethical risks of technologies deployed in society will be a key focus of future research.

Third, the Lack of Value Trade-off and Decision-Making Frameworks. When fundamental conflicts arise between "accuracy" and "fairness," "efficiency" and "privacy," or "safety" and "autonomy," what framework should engineers use to make decisions? This is essentially a value judgment, yet it is an unavoidable daily issue in engineering practice. For example, to improve the safety of an autonomous driving system (protecting pedestrians), is it acceptable to sacrifice some passenger privacy (through more intensive in-car monitoring)? Who should have the authority to make this decision? The engineers, the company, regulators, or the public? EtEn cannot merely provide a set of potentially conflicting tools; it must develop structured, procedural frameworks for trade-offs and decision-making. This may include consensus-based prioritization, public participation mechanisms based on democratic deliberation, or clear legal rules. The absence of such frameworks places engineers under tremendous moral and professional risk.

Fourth, the Challenge of Auditability and Accountability. Whether a system is ethical cannot be determined solely by the developers' self-certification; it must be verified through independent, repeatable audits. However, there is currently a lack of widely accepted algorithmic audit standards, tools, and professional audit teams. Technical black boxes (especially deep learning models) make auditing exceptionally difficult. EtEn must promote the development of XAI tools and use them as the foundation for ethical audits. Simultaneously, legal accountability needs clarification: When an accident

occurs, how is responsibility allocated among designers, developers, deployers, users, and even the algorithm itself? Establishing a clear chain from technical traceability to legal accountability is the institutional guarantee for the implementation of EtEn.

Fifth, the Boundaries and Risks of Standardization. Standardization is a cornerstone of engineering, yet the development of ethical standards presents a distinct double-edged challenge. While providing crucial guidance and stability for industry adoption, an overtly rigid or minimalist approach to standardization risks fostering "checkbox-ticking compliance," where meeting minimum requirements becomes the endpoint, inadvertently stifling moral imagination and superior ethical practices that exceed baseline norm. To navigate this tension, EtEn must actively advocate for and contribute to the design of dynamic, process-oriented standards. These should function less as static checklists and more as evolving frameworks that mandate continuous improvement.

Sixth, Cross-Cultural Global Governance Coordination. Technology is borderless, whereas values are regional. The differing ideas of China, the US, and Europe regarding data privacy, freedom of speech, and social governance have led to divergent paths in technological governance. The development of EtEn must confront the challenge of building global governance coordination. A completely fragmented governance system could lead to "ethical protectionism" and "regulatory arbitrage" (i.e., developing and testing in jurisdictions with the loosest standards), while imposing uniformity would ignore legitimate cultural diversity. Therefore, the key challenge for the future is how to form global consensus on core issues that prohibit a "race to the bottom" (e. g., lethal autonomous weapons), while exploring cooperative mechanisms like mutual recognition of certifications and cross-border data flows in other areas, seeking minimal consensus based on respect for diversity.

Seventh, Advanced AI Alignment and the Risk of Ethical Loss of Control. Facing the potential future emergence of superintelligence (AGI, Artificial General Intelligence), EtEn encounters its ultimate challenge — the Alignment Problem: How to ensure that the ultimate goals of an AI system far surpassing human capabilities remain fully aligned with humanity's complex, ambiguous, and dynamically evolving values? (Ji et al., 2023) This goes far beyond the current scope of algorithmic fairness, involving the internal modeling and calibration of motives, intentions, and values. Failure could be existential. EtEn must begin to prospectively consider these "long-term future" problems.

Eighth, The Ethics of Ethical Engineering Itself. Finally, we must maintain critical reflection on EtEn itself. This powerful "hammer" could also be misused. Who decides which ethics are to be "materialized"? Could it become a tool of techno-authoritarianism, enabling social control and value indoctrination through technological design under the guise of "being good for you"? Therefore, the development process of EtEn itself must be transparent, democratic, and responsible. It must incorporate a mechanism for self-criticism and self-correction, remaining vigilant against the risks of alienation it may introduce, and ensuring that it ultimately serves human well-being and social prosperity, not the specific interests of any single group.

These eight key challenges are both severe tests and defining opportunities for EtEn. They necessitate the development of what may be termed "hybrid knowledge," a deeply integrated, interdisciplinary framework where philosophers and ethicists clarify normative foundations, social scientists map contextualized value interpretations, and engineers co-develop corresponding formalisms and technical implementations— especially translating abstract ethical principles into concrete technical realizations through algorithmic formalization and system design. Only through such cross-boundary concerted efforts can EtEn overcome its numerous obstacles and evolve from a promising concept into a mature discipline and practical system capable of reliably guiding the course of technological development and ensuring it benefits humanity. The success of this endeavor concerns not only the fate of one discipline but the future of us all.

5. CONCLUSION

The growing complexity and pervasiveness of emerging technologies underscore the urgency of embedding values such as fairness, accountability, and transparency into engineering practice. Yet significant challenges remain, including a shortage of practical tools, limited interdisciplinary collaboration, and underdeveloped methodologies for ethical evaluation. These gaps highlight the need for a coordinated international effort to advance EtEn as a novel engineering discipline. The transition from EnEt to EtEn represents a necessary evolution in the governance of emerging technologies. While educating ethical practitioners remains essential, it is no longer sufficient. This shift demands systematic approaches that integrate ethical considerations directly into technological design and development.

AI is central to this transformation. Tools such as XAI, ethics-embedded algorithms, and predictive risk models enable the operationalization of ethics, making governance more scalable and proactive. However, technical

solutions alone cannot address persistent challenges such as value alignment and algorithmic bias. A holistic approach is required — one that aligns with Fisher's (Fisher, 2019) concept of "socio-technical governance," which merges technical capabilities with robust social oversight, transcending purely technical or exclusively social models. Crucially, the vision of "using AI to govern AI" must be implemented within a human-in-the-loop framework, where automation augments rather than replaces democratic deliberation and human judgement. Ultimately, the goal is to couple advanced technological tools with deepened democratic engagement, forming an adaptive and reflective EtEn capable of addressing the ethical, legal, and social implications of emerging technologies. This paradigm not only expands practical pathways for implementing EnEt but also promotes responsible innovation in the age of AI.

It should be emphasized that EtEn is not a panacea but a socio-technical approach whose success depends on addressing its limitations through continued technical refinement and robust human oversight, and that effective EtEn depends on sustained collaboration among educators, researchers, engineers, policymakers, and civil society. Philosophers and ethicists should play a key role in propelling the development of ethical engineering rather than remaining in an ivory tower of pure discourse. Only through shared commitment and cross-sectoral dialogue can we ensure that technological advancement serves humanity's best interests, promoting not only economic growth but also equity, dignity, and collective well-being for generations to come.

REFERENCES

Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6:52138–52160.

Adobor, H., and R. Yawson. 2023. "The Promise of Artificial Intelligence in Combating Public Corruption in the Emerging Economies: A Conceptual Framework." *Science and Public Policy* 50:355–370.

Brey, P. 2012. "Anticipatory Ethics for Emerging Technologies." *NanoEthics* 6 (1): 1–13.

Camacho, N. G. 2024. "The Role of AI in Cybersecurity: Addressing Threats in the Digital Age." *Journal of Artificial Intelligence General Science* 3 (1): 143–154.

Collin, B., P. Izmailov, J. H. Kirchner, et al. 2023. "Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision." arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2312.09390.

Dwivedi, R., D. Dave, H. Naik, et al. 2023. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions." *ACM Computing Surveys* 55 (9).

Elliot, N., E. Katz, and R. Lynch. 1993. "The Challenger Tragedy: A Case Study in Organizational Communication and Professional Ethics." *Business & Professional Ethics Journal* 12 (2): 91–108.

Feigenbaum, A. V. 2012. *Total Quality Control*. 4th ed. New York: McGraw-Hill.

Fisher, E. 2019. "Governing with Ambivalence: The Tentative Origins of Socio-Technical Integration." *Research Policy* 48 (5): 1138–1149.

Friedman, B., and D. G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge (MA): The MIT Press.

Friedman, B., P. H. Kahn, A. Borning, and A. Huldtgren. 2013. "Value Sensitive Design and Information Systems." In *Early Engagement and New Technologies : Opening up the Laboratory*, ed. by N. Doorn, D. Schuurbiers, I. van de Poel, and M. Gorman, 55–95. Dordrecht: Springer.

Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds & Machines* 30 (3): 411–437.

Gou, Z., Z. Shao, Y. Gong, et al. 2023. "CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing." arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2305.11738.

Harris Jr., C. E., M. S. Pritchard, and M. J. Rabins. 2013. *Engineering Ethics: Concepts and Cases*. Wadsworth: Cengage Learning.

Hersh, M., ed. 2015. *Ethical Engineering for International Development and Environmental Sustainability*. London: Springer.

Jain, V., A. Balakrishnan, D. Beeram, et al. 2024. "Leveraging Artificial Intelligence for Enhancing Regulatory Compliance in the Financial Sector." *International Journal of Computer Trends and Technology* 72 (5): 124–140.

Ji, J., T. Qiu, B. Chen, et al. 2023. "AI Alignment: A Comprehensive Survey." arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2310.19852.

Madaan, A., N. Tandon, P. Gupta, et al. 2023. "Self–Refine: Iterative Refinement with Self-Feedback." arXiv. Accessed June 1, 2025. https://arxiv.org/abs/2303.17651.

Martin, M. W., and R. Schinzinger. 2010. *Introduction to Engineering Ethics*. New York: McGraw-Hill.

Padmanaban, H. 2024. "Revolutionizing Regulatory Reporting through AI/ML: Approaches for Enhanced Compliance and Efficiency." *Journal of Artificial Intelligence General Science* 2 (1): 57–76.

Schlossberger, E. 2023. *Ethical Engineering: A Practical Guide with Case Studies*. Boca Raton: CRC Press.

Ulnicane, I., T. Mahfoud, A. Salles, et al. 2023. "Experimentation, Learning, and Dialogue: An RRI-Inspired Approach to Dual-Use of Concern." *Journal of Responsible Innovation* 10 (1).

Umbrello, S., and I. van de Poel. 2021. "Mapping Value Sensitive Design onto AI for Social Good Principles." *AI and Ethics* 1:283–296.

Urquhart, L. D., and P. J. Craigon. 2021. "The Moral-IT Deck: A Tool for Ethics by Design." *Journal of Responsible Innovation* 8 (1): 94–126.

Van de Poel, I. 2020. "Design for Value Change." *Ethics and Information Technology* 23 (1): 27–31.

Verbeek, P.-P. 2006. "Materialising Morality: Designing Ethics and Technological Mediation." *Science, Technology and Human Values* 31:361–380.

——— . 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.

Wallach, W., and C. Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wang, D. 2018. "Toward an Experimental Philosophy of Engineering." In *Philosophy of Engineering* : *East and West*, ed. by C. Mitcham, B. Li, B. Newberry, and B. Zhang, 37–50. Berlin and Heidelberg: Springer.

——— . 2020. "Towards Responsible Engineering: Interpretation and Implementation of Ethical Codes." *Chemical Engineering Higher Education* 37 (3): 1–7.

——— . 2023. "Towards Ethical Engineering: Artificial Intelligence as an Ethical Governance Tool for Emerging Technologies." *Computer Sciences and Mathematics Forum* 8.

Wang D. [Ван Д.] From Engineering Ethics to Ethical Engineering [От инженерной этики к этической инженерии] : Leveraging AI for Governing Emerging Technologies [использование ИИ для управления перспективными технологиями] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 47–67.

Дачжоу Ван
д. филос. н., профессор, Школа гуманитарных наук Университета Китайской академии наук (Пекин); ORCID: 0000–0001–9586–4597

От инженерной этики к этической инженерии

использование ИИ для управления перспективными технологиями

Получено: 16.09.2025. **Рецензировано:** 11.11.2025. **Принято:** 11.11.2025.

Аннотация: Этическая инженерия (ЭтИн) — это формирующаяся дисциплина, представляющая собой смену парадигмы по сравнению с традиционной инженерной этикой (ИнЭт). В отличие от подхода, ориентированного в первую очередь на обучение отдельных специалистов, ЭтИн ставит целью системное внедрение этических принципов в саму структуру технологических систем и процессов управления. В статье анализируется переход от ИнЭт, ориентированной на обучение специалистов нормативным принципам, к ЭтИн, рассматривающей этику как инженерную задачу системного уровня, предполагающую трансляцию этических принципов в исполняемые инструменты управления. В исследовании подчеркивается двойственная роль ИИ как и основной области, требующей регулирования, и ключевого средства для его реализации, а также анализируется

его потенциал для совершенствования этического управления через повышение проверяемости алгоритмов, поддержку сложного этического принятия решений и междисциплинарное коллаборативное управление. Одновременно рассматриваются такие вызовы, как обеспечение согласования ценностей, снижение рисков предвзятости и опасность технологического редукционизма. Работа выделяет восемь ключевых проблемных областей, формирующих ядро исследовательской повестки ЭтИн, и утверждает, что развитие этой дисциплины следует понимать как экспериментальный, итеративный процесс. Сформулированный парадигмальный сдвиг не только расширяет практические возможности реализации инженерной этики, но и предлагает новые методологические решения для этического управления развивающимися технологиями в эпоху ИИ.

Ключевые слова: инженерная этика, этическая инженерия, искусственный интеллект (ИИ), перспективные технологии, ценностно-ориентированное проектирование, морализирующие технологии.

DOI: 10.17323/2587-8719-2025-4-47-67.

Grunwald, A. 2025. "Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 68–83.

Armin Grunwald*

Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions**

Submitted: Aug. 13, 2025. **Reviewed:** Oct. 19, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This paper deals with the possibility of gradual disruptions at the societal level in the course of rapidly advancing digitalization and spread of AI. The term "disruption" refers to the sudden breakdown of familiar, previously stable constellations. Expectations of stability, assumptions of continuity, and planning security are shattered, casting the prospects for the future in an uncertain light. The Latin roots of the term mean "bursting," "breaking," and "tearing," semantically referring to the temporal structure of more or less sudden, abrupt events. Seen in this light, the talk of gradual disruption in the title of this article seems conceptually contradictory or paradoxical. However, there are many examples of disruption in the world of technology that were heralded by recognizable but often unnoticed signs, particularly by material fatigue and wear. The daily stress on many technical objects, such as V-belts in older vehicles or bridge structures, gradually leads to wear and degradation. In this sense, the notion of gradual disruption refers to upheavals with significant or even dramatic damage potential that do not occur unexpectedly and suddenly, like a global pandemic or an earthquake, but build up gradually until they finally lead to the disruption of previously stable constellations. I will argue that this type of potential and gradual disruption could emerge in areas of digitalization and AI. Examples include the increasing but unnoticed standardization of human actions, the silent loss of freedom and individuality, the increasing dependence on the smooth functioning of digital infrastructures, the loss of the future as an open space, or the loss of reflection and learning opportunities due to unlimited acceleration. The possibility of such gradual disruptions poses several challenges to responsible research and innovation (RRI), technology assessment (TA), and ethics. These include epistemological issues (how to detect gradual disruptions at an early stage), ethical issues (how to assess and evaluate concerns relating to the precautionary principle, for example), issues of whether countermeasures should be taken, and issues of communication between irrational exaggeration and irrational trivialization. The final part of the paper will address possible gradual disruptions that can be attributed to both technical parameters and human behavior, and draw conclusions for TA and RRI.

Keywords: Disruption, Digital Twin, Technological Dependence, Loss of the Future.

DOI: 10.17323/2587-8719-2025-4-68-83.

*Armin Grunwald, Dr. rer. nat., Dr. habil. in Philosophy (venia legendi); Professor at the Institute of Philosophy, KIT (Karlsruhe, Germany); Head of the Institute for Technology Assessment and Systems Analysis (Karlsruhe, Germany); Head of the Office of Technology Assessment at the German Bundestag (Karlsruhe, Germany), armin.grunwald@kit.edu, ORCID: 0000–0003–3683–275X.

1. INTRODUCTION AND OVERVIEW

Since the Second World War and especially in recent decades, technological progress has become a key factor in social development in many areas. Innovation and competitiveness require new technologies, such as in digitalization, medicine, or biotechnology, as well as for the transition to a more sustainable and climate-friendly society. However, this has led not only to the desired consequences but also to unintended, sometimes surprising, and often undesirable and problematic ones (Grunwald, 2019). These include major accidents in technical facilities (e.g., Bhopal and Chernobyl), the global environmental crisis (e.g., loss of biodiversity and climate change), stress for the labor market due to automation, risks for democracy due to problematic internet communication as well as the potential for dual use and misuse of technology at various levels. In current times, the divergence between intended consequences of technology and innovation and unintended ones, often manifesting themselves years or decades later, coincides with the emergence of a multi-polar world full of geopolitical tensions, including political competition in major areas of new technology like AI, robotics, quantum technologies, and biotech.

In this situation, *forward-looking* analysis and assessment of technology impacts are essential, in terms of both opportunities and possible unintended negative consequences. This diagnosis inspired the introduction of technology assessment (TA) in the US Congress in 1972 as scientific policy advice on the design and impact of technology (Bimber, 1996). This was followed by the spread and diversification of TA. Three main fields of practice can be distinguished today (Grunwald, 2019):

- TA as scientific policy advice, e.g., at the German Bundestag (see below), addresses publicly relevant, generally binding aspects of technology that must be decided by policy-makers, such as safety and environmental standards, the protection of citizens, the guarantee of human and civil rights, or the priority setting in research funding and technology policy.
- TA to support public debate and opinion-forming systematically engages citizens and stakeholders in debates on future technology, frequently involves the mass media, and sees itself as an element of deliberative democracy at grassroots level, beyond the institutions of representative democracy.
- TA in direct technology design accompanies the research and development of technology at universities and in industry. TA's knowledge

of consequences is incorporated directly into the development of technology, e.g., in order to design technology in line with values such as sustainable development or privacy.

Technological consequences are not simply the consequences of technology. They depend not only on technical parameters but also arise from the interaction of technical properties and human behavior, for example, through use and consumption. TA is therefore ultimately not about technology as such, but about researching and shaping socio-technical interactions. For this reason, TA is necessarily highly interdisciplinary and must involve engineering, social sciences, and ethics in particular. This applies equally to the field of *responsible research and innovation* (RRI; Von Schomberg & Hankins, eds., 2019).

Unintended consequences of the digital transformation have a different character than those of many other technologies. While TA has often had to deal with environmental, health, or accident risks in its history, for example, in the context of nuclear energy, these types of risks do not play a central role in digitalization. Instead, fears are repeatedly expressed here that can be understood as concerns about *gradual disruptions at a societal level*, i.e., about slow developments that can nevertheless grow into upheavals with considerable potential for damage (Section 2). Such possible upheavals characterize the debate on digitalization (Section 3).[1] They pose specific challenges for TA and RRI (Section 4).

2. ON THE CONCEPT OF GRADUAL DISRUPTION

Disruption has only become a widely used term in the last ten years or so. Although the word's origin refers to rather unpleasant-sounding meanings (lt. *disrumpere*, "to burst, break, tear apart"), it entered contemporary discourse with a positive intention. Disruptive innovations, based on technological leaps or entirely new business models, are valued in innovation policy (Vera & Ramge, 2021). In contrast to incremental innovations based on gradual product improvements, disruption is aimed at fundamental upheaval intended to overturn market conditions that have existed for years or even decades within a short space of time. New market opportunities are then open to the winners (often called "disruptors"). Also, entirely new markets can emerge, as in the digital transformation exemplified by platform economies such as Amazon or eBay or in digital photography.

[1] This publication continues and deepens earlier work by the author (cp. Grunwald, 2025).

In this way, the historically older theory of disruptive technology (Bower & Christensen, 1995) was quickly extended to the field of disruptive innovation (e.g., Danneels, 2004), in some cases with considerable expectations. However, the term is controversial (Gans, 2017): "'Disruption' is a business buzzword that has gotten out of control. Today everything and everyone seem to be characterized as disruptive — or, if they aren't disruptive yet, it's only a matter of time before they become so." In this criticism, the concept of disruption is reduced to a synonym for success.

For some years now, *crisis phenomena* have also been referred to as disruption. The coronavirus pandemic and recent geopolitical tensions are considered disruptive events. Both have ended a long period of broad stability, at least in the Global North, and, according to widespread diagnosis, indicate the transition to a time of permanent crisis. The term disruption is used here to describe the breakdown of stable social conditions. In communication, catastrophic narratives often come into play, such as the fear of nuclear war, climate change as the end of the Earth's habitability, the end of democracy, or the collapse of the labor market due to massive automation. Expectations of stability, assumptions of continuity, and planning certainty are breaking down and making future prospects appear uncertain. Semantically, this points to the time structure of abruptly occurring events. Seen in this light, talk of *gradual* disruption seems conceptually absurd or paradoxical.

A closer look allows us to differentiate. Semantically, the term disruption shows two elements of meaning: on the one hand, the *breakdown* of previously stable relationships and, on the other, the *speed of* this breakdown. While the first element of meaning is etymologically inherent in the term, the second can be handled more flexibly. Time scales of breakdown are elastic. For example, the invention of printing in the late European Middle Ages is often portrayed as disruptive — historically, this disruption extended over many decades of diffusion into the societies of the time. So, *on the one hand*, breakdown and discontinuation can certainly take place over an extended period of time. They are then only referred to as disruption in retrospect, in a kind of fast motion so to speak, whereas they appeared to the contemporary witnesses as a gradual transformation. *On the other hand*, discontinuation and breakdown, even if they occur suddenly, can build up slowly over longer periods of time. Nevertheless, everything remains stable for a long time and disruption only occurs later. The latter are referred to as *incremental* or *gradual disruptions*. This is what this article is about.

Many examples of this type of disruption are known from the technical world, especially those involving material fatigue and wear. The daily stress on many technical objects, such as bridges or V-belts in cars, gradually leads to the degradation of materials and components. They still function reliably for a long time until the wear and tear reach a level where something fails from one moment to the next, so that, in the example chosen, the V-belt suddenly breaks or the bridge collapses without warning, as happened in Dresden in 2024. In the medical field, strokes and heart attacks fall into this category. Certain signs can be recognized in advance with some uncertainty, such as calcium deposits in arteries, but the event then happens suddenly and unexpectedly. People often ask afterwards whether one could have known about it beforehand. One example from the climate debate is the so-called *tipping points* (see Gladwell, 2000). Further warming could lead to self-reinforcing feedback effects that would have dramatic consequences in a short space of time, i.e., a disruptive effect.

The disruptive effect in processes of this kind is therefore inherent in incremental processes that are difficult to recognize. It can remain unrecognized for a long time and thus escape early intervention and prevention. At some point, however, it can lead to potentially far-reaching and sudden consequences. The tragedy of such gradual developments is that in the incremental course, serious disruptions may announce themselves gradually, but can then take place abruptly. With this semantic differentiation, the possibility *of gradual disruption* in digitalization and the mass introduction of AI is considered in the following.

3. DISRUPTIVE POTENTIAL OF TRANSFORMATION THROUGH AI

The term *gradual disruption* can be used analytically to address possible developments in the digital and AI transformation with damaging or even catastrophic potential. This is not about predictions but about *possible* developments and corresponding concerns. They can be found at different levels in the debates on AI and digitalization. The following series of examples does not follow an ordering principle and does not claim to be exhaustive but reflects facets of the current debate on AI and digitalization (e.g., Deutscher Ethikrat, 2023).

SLIPPING INTO DEPENDENCIES

Modern societies are already completely dependent on the smooth functioning of critical infrastructures such as the power supply (Petermann et al., 2011). This increasingly applies to digital infrastructures. If the internet

were to fail, financial transactions would become impossible, the global economy would collapse, media communication would no longer be possible, medical diagnostics would be deprived of many established procedures, international logistics chains would come to a standstill, and much more. The increasing introduction of ADM (*automated decision-making*) systems is creating a dependency on AI-controlled systems, which, together with their *black box* character and lack of transparency, but also due to the psychological *automation bias* (Deutscher Ethikrat, 2023; Safdar et al., 2020), lead to increasing dependence on these systems in decision-relevant contexts such as the police and social services.

The gradual displacement of cash is a current example of the ambivalence of technical infrastructures. While cashless payment transactions were initially an *additional* option alongside cash as a convenience for businesses and private individuals, there is now a gradual transition to a world without cash (Ehrenberg-Silies et al., 2022). Cash is slowly being displaced, partly driven by consumer behavior and convenience, partly driven by political and economic incentives and regulation, with the argument that this could make the black market and illegal work impossible. Once cashless payment transactions have become fully established, as is already largely the case in some countries, the freedom of choice in payment options will have disappeared and, if the internet were to go down, no more shopping or payment transactions would be possible. Due to its gradually increasing dominance, cashless payment becomes a compulsion, accompanied by full dependence on technical systems.

Dependencies are not disruptions in themselves, but they carry their seeds. Dependencies that have become total are *latent disruptions*. As disruptions in waiting, they build up gradually through growing dependencies, but in an emergency, e.g., if digital technologies were no longer to function smoothly, they can have abrupt and possibly catastrophic consequences. However, relying on their unlimited smooth functioning and making the functionality and stability of modern societies dependent on it is ethically problematic. Unexpected hacker events, a collapse of state order, or severe economic turmoil could also affect infrastructures such as the internet and, in the worst case, render them dysfunctional. Complete dependence on digital infrastructures and platforms is likely to have been reached long ago — which means that modern societies are already operating in the mode of this latent disruption.

LOSS OF THE FUTURE AS AN OPEN SPACE OF POSSIBILITIES

Digital technologies are often regarded as synonymous with the future, much like nuclear energy optimistically was in the so-called atomic age of the 1950s and 1960s. However, digital technologies generally operate based on past data. For example, digital twins (see above) only ever mirror a world of yesterday, e.g., in that customer profiles can only be created based on past purchase and consumption processes. Digital twins basically only depict the *past* of their analog originals. Big data technologies can only evaluate past data and recognize past patterns. AI systems can only be trained on data from the past, as data from the future is not available. Even if AI and *big data* are used to create quantitative forecasts, these are based on pattern recognition using past data. Due to the indispensable reference to data, digital technology is inescapably fixated on past conditions. When data sets, digital twins, and correlations and patterns uncovered by AI are used to make predictions about the future, past conditions are carried over into the future, imposed on it, so to speak. The future as an at least partially open space of alternative paths and possibilities is replaced by a data-based extension of the past.

Digitalization or some of its fields could become conservative in this way, aligning concepts for the future with old data rather than developing new ideas. Multiple anthropological determinations understand humans as beings with a future and the ability to envision and reflect on possible futures (e.g., Kamlah, 1973) — futures that go beyond extending the past into to the future instead include creative ideas in an open space of possibilities, which may even have a counterfactual and utopian character. A gradual disruption could occur here if the fundamental openness of the future fades into the background or disappears completely in favor of a data-driven orientation that remains bound to the past.

GRADUAL DISAPPEARANCE OF FREEDOMS

Human freedom again and again leads to unwanted effects. The example of road traffic, with over one million deaths worldwide every year, the majority of which are due to human error, is one example; crime and terror are others. Promises of security through prevention of accidents or defense against terrorism repeatedly provide arguments for interfering with civil liberties through surveillance and control. Regulation, the legal system, and security agencies should ensure that people do not exercise their freedoms at the expense of others. Technical surveillance and control systems are

used to technically enhance security or to enforce it completely. Digitalization provides powerful tools for this (Spiekermann & Christl, 2016). Comprehensive surveillance by cameras, automated facial recognition, location tracking and creation of movement profiles, pattern recognition in offender profiles, technical requirements in operation, and even the removal of the "human factor" from technical processes, such as in autonomous driving, offer far-reaching opportunities to technically prevent the abuse of human freedoms — but also to abolish freedoms. Quite a few countries, especially in Southeast Asia, have achieved a high degree of technically implemented security and control in this way.

However, if security is enforced through technical means, there is no longer any freedom in the field concerned (Deutscher Ethikrat, 2023: 357). The tension between the high value placed on freedom, rooted in the European Enlightenment, on the one hand, and technical control and surveillance, e.g., to ward off terrorism, on the other, is a recurring theme in social debates on digital transformation.

The gradual disruption in this field would be an unnoticed slide into a world where the security interests of individuals and the state become the dominant value and are no longer weighed against other values, such as civil liberties. This would lead to ever-increasing control of human actions enforced by digital technology. Such a development would spell the end of individual freedom and erode the democracy based on it. The result would be a society controlled by digital means that is secure but completely unfree. Science fiction has repeatedly addressed such dystopian developments.

DIFFUSION OF RESPONSIBILITY INTO NOWHERE

Responsibilities are being redistributed at the constantly emerging new interfaces between humans and digital systems. Automated or autonomous decision-making systems (ADM systems), industrial production in cooperation between humans and robots in Industry 4.0, and autonomous driving are examples. However, the fact that machine systems are responsible for certain decisions does not mean that these systems also bear responsibility. This is because even AI-supported systems do not have intentions but merely use algorithms to perform complex statistical operations based on data. If they do not follow their own agenda and do not consciously want to achieve a specific purpose, they cannot bear responsibility (ibid.). The attribution of responsibility and accountability remains, at least for the time being and in the foreseeable future, the preserve of humans who act consciously and with intentions.

However, the attribution of responsibility to specific actors is becoming increasingly complex in a world with more and more AI systems. Although decisions and therefore responsibility in principle remain with humans, this is increasingly happening invisibly. While in a traditional car, the person driving is obviously responsible, this is much harder to recognize in autonomous cars. AI systems and their manufacturers interpose themselves between the intentionally acting humans and real effects, e.g., in the event of a traffic accident caused by an autonomous car. Responsibility shifts from individual drivers or, in the case of military drones, from soldiers to people and institutions in the background — to companies, programmers, managers, secret services, generals, or regulatory authorities.

Ethics and law have experience in assigning responsibility in complex contexts involving a division of labor, e.g., in large companies. The task of defining responsibility in constellations based on the division of labor between humans and AI systems also appears to be feasible in principle. However, the complexity of responsibilities distributed between humans and AI systems increases both the risk of a gradual "diffusion of responsibility" into nowhere and the risk of deliberate concealment of responsibility. In light of the philosophical ideals of linking freedom with responsibility, it is completely open whether and what kinds of possible gradual disruptions in the social order may result.

THE GRADUAL UNLEARNING OF ESSENTIAL SKILLS (DESKILLING)

In the course of historical and technological change, there are always skills that become dispensable and are forgotten. Examples of past professions that are no longer needed today can be found in museums. Knowledge is also lost, prompting historians and archaeologists to ask questions such as how the pyramids in Egypt or the Gothic cathedrals could have been built with the technical means available at the time. Such processes of forgetting take place slowly; new skills are developed in place of those that have been forgotten.

Digitalization, automation, and AI lead to similar, but massively amplified, more far-reaching, and accelerated effects. They make life pleasant and convenient in many ways, relieve people of the need to orient themselves in space through GPS, render learning superfluous in many fields, as knowledge is available digitally, and replace lengthy deliberations and the need to form an independent judgment through data-based calculations in ADM systems. Many fear that this could lead to the atrophy of abilities that are *constitutive* of being human (*deskilling*; cf. Deutscher Ethikrat, 2023: 353). As a result of

the widespread use of AI applications, people could be increasingly tempted to delegate more and more tasks to AI technology because it is seen as supposedly superior or because it is convenient and saves time and effort.

An important role is played here by a psychological effect that is specific to AI systems: the *automation bias* (cf. Goddard et al., 2014). Many people trust algorithmically generated results based on large amounts of data and calculated using AI-supported decision-making processes more than those of human experts, no matter how much professional and life experience they have. The reason for this presumably lies in exaggerated attributions of objectivity and accuracy toward mathematical and data-based processes, on the one hand, and a suspicion of inaccuracy and subjectivity toward humans, on the other. Even if AI systems are strictly limited to *decision support* and human decision-makers have to make the decision, AI systems could gradually take on the role of the "actual" decision-makers and thus substantially erode human judgment.

In this way, key human skills and cultural techniques could be pushed into the background and eventually atrophy. Examples include the ability to understand complex texts when relying solely on short summaries generated by ChatGPT, or the ability to form one's own opinion on a complex issue when permanently and uncritically relying on the decision support of an AI application. The gradual disruption here would be the combination of the loss of human abilities such as judgment and critical thinking with an increasing dependence on AI systems.

END OF OPPORTUNITIES FOR REFLECTION AND LEARNING

Acceleration is part of the capitalist economic system. It unleashes creativity and innovation, primarily through competition. Acceleration is a phenomenon often discussed in the context of digitalization. The increase in computing speed, the possibility of calculating millions of options in the shortest time, the linking of creative resources via the internet, and the acceleration of data transfer and communication, much of it mediated and further accelerated by means of digital twins, all shorten innovation cycles. Accordingly, the above-mentioned "disruptive innovation" as extreme acceleration is the opposite of incremental innovation processes.

However, there is also destructive competition. The acceleration spiral is in danger of overexploiting the human and natural resources that feed it. One concern regarding AI-driven digitalization relates to negative and potentially ruinous consequences of ever-increasing acceleration, in particular to the question of whether and when further acceleration could fundamentally

undermine important conditions of reflection. This would be contrary to the principles of enlightenment, the principles of technology assessment (Grunwald, ed., 2024), and the requirements of sustainable development.

Reflection requires careful analysis and deliberation, weighing alternatives, finding the right balance and ethically legitimate criteria for decision-making, as well as prudent implementation of the results, e.g., in legal regulation. All of this takes time in two ways: first, for the deliberation and consideration processes themselves and, second, for transferring the results into practical action and decision-making. The gradual disruption in this respect could be that societal capacities and structures for reflection would slowly erode under the pressure of capitalist competition. In the libertarian narrative of an innovation-oriented fatalism under the primacy of competitive thinking, reflection can no longer be afforded, since otherwise the competitor would be faster and gain market advantages.

4. REQUIREMENTS FOR TECHNOLOGY ASSESSMENT

Some of the developments described have already taken place (complete dependence on digital systems), some are observable (unlearning of skills), and some are only feared (diffusion of responsibility into nowhere). None of them are consequences of technology alone. Rather, gradual disruptions in connection with AI and digitalization arise from a combination of technical possibilities, business models, human behavior, and regulations. For example, AI does not actively take over the thinking for people, but people give up thinking for themselves and let "AI do the thinking." Another example: dependence on AI systems arises from the fact that almost all routines in business and politics, but also in leisure and everyday life, now run via digital channels, and many are supported by AI. This is not a predetermined consequence of the existence of AI, but individuals, communities, or entire societies allow themselves to become so accustomed to these technologies in their behavior and habits that they are gradually becoming dependent. Digital technology and AI, together with applications and business models, provide the medium for gradual disruption, but are not solely responsible for it.

Digital and AI systems offer so many advantages that there is a pull toward their use and adaptation. In the process, digital systems are often overly trusted and people risk losing their own expertise (automation bias, see above). Undoubtedly, digital technology often makes life pleasant and convenient. As soon as routine activities at work or during leisure time have been adapted to digital systems, whether with or without AI, they are so much

a part of life that it is often hard to imagine life without them, or at least it seems tedious and exhausting, hence unattractive. Such effects can also result from time pressure and efficiency requirements at work. If, for example, the individual examination of the records for processing an application for "Bürgergeld" (citizens' income) requires working through extensive documents and takes time, while algorithms can do this quickly, provided the data is digitized, the willingness to let the algorithms do the work increases.

So it is not technology as such that leads to gradual disruption, but its combination with human behavior. When considering the consequences, the focus must therefore not be narrowed down to digitalization and AI as technology, but must instead take into account the interactions with human behavior. For TA, this is a fundamentally familiar but comparatively difficult constellation. The often only vaguely tangible human factor in terms of convenience, adaptation, and overestimation of digital systems — perhaps most strongly the "sweet temptation" of convenience — adds to the usual difficulties in recognizing gradual processes and assessing their relevance for action.

Gradual, creeping developments are often difficult to recognize at first. This is particularly true in their early phases, when insufficient data is available and only weak signals can be observed. The weak evidentiary basis, the lack of sensitivity to the only slowly developing potential for disruption, and the uncertainty as to whether a disruptive development will occur at all often reduce the willingness to deal with these developments proactively and, for example, to conduct empirical research to examine the evidence. Only when the signs of a disruptive development become more apparent does this willingness increase. Climate change as a structurally analogous, gradual disruption has provided illustrative material on scientific uncertainty and the growth of evidence since the 1970s. In digitalization, concerns about democracy were also initially rather speculative (Grunwald et al., 2006), whereas they have long since been empirically proven (Hofstetter, 2016). Also, the loss of competence due to the transfer of tasks to digital systems is no longer just a fear but has been substantiated by many examples in connection with the "ironies of automation" (Bainbridge, 1983).

In TA, the epistemological complexity is well known from conflicts over precaution, particularly in the health and environmental fields (Harremoes et al., eds., 2002). In typical precautionary situations, there is little knowledge about potential future damage and its probability of occurrence (Jonas, 1979; Von Schomberg, 2005). This epistemological challenge has direct consequences for the assessment and classification of developments that

are only gradually becoming visible. The conclusion that TA and ethics should hold back until better knowledge is available (Nordmann, 2007) is out of the question in view of the high relevance of potential disruptions, in particular because of possible *points of no return*. However, prioritizations and urgency assessments require a certain level of evidence of knowledge about potential disruptions (Grunwald, 2010). A mere suspicion is not sufficient for a high prioritization, even if it would lead to a disastrous development if the suspicion were confirmed. There is the difficult task of assessing the situation and classifying it in comparison with other developments. The question arises as to when the evidence of a suspicion is sufficiently strong to at least legitimize the allocation of resources for more research in this area or even for intervening measures for preventive counteraction (Von Schomberg, 2005).

Due to the poor recognizability of gradual developments and the difficulties in assessing them, public communication about them is susceptible to ideology and speculation. On the one hand, there is a tendency to trivialize and downplay the issue, arguing that one should wait until better data and corroborated evidence is available instead of rashly wasting resources or unnecessarily restricting freedoms. On the other hand, weak signals are extrapolated into the future and dramatized to the point of stoking fears of rapid disruption. This results in mutual accusations of exaggeration, ideology, speculation, trivialization and whitewashing, as well as recklessness, irresponsibility, or permanent doubting. During the corona pandemic, these communication problems could be observed in many ways. Time and again, there seemed to be no path of mediating reason between dramatizing exaggeration on the one hand and downplaying the risks on the other.

Given the wide differences in the perception of opportunities and risks, there is certainly no one-size-fits-all solution to these communicative challenges. However, past debates on technology (Grunwald, 2011) show that neither trivialization nor dramatization are appropriate communication patterns. What is constructive is transparency and openness, including, and perhaps especially, with regard to the uncertainties of knowledge and the possible extent of damage.

REFERENCES

Bainbridge, L. 1983. "Ironies of Automatization." *Automatica* 19 (6): 775–779.

Bimber, B. A. 1996. *The Politics of Expertise in Congress: The Rise and Fall of the Office of Technology Assessment*. New York: State University of New York Press.

Bower, J. L., and C. M. Christensen. 1995. "Disruptive Technologies. Catching the Wave." *Harvard Business Review* 69:19–45.

Danneels, E. 2004. "Disruptive Technology Reconsidered. A Critique and Research Agenda." *Journal of Product Innovation Management* 21 (4): 246–258.

Deutscher Ethikrat. 2023. *Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz* [in German]. Berlin: Deutscher Ethikrat.

Ehrenberg-Silies, S., R. Peters, C. Wehrmann, and S. Christmann-Budian. 2022. *Welt ohne Bargeld — Veränderungen der klassischen Banken- und Bezahlsysteme* [in German]. Berlin: Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag.

Gans, J. 2017. *The Disruption Dilemma*. Cambridge (MA): The MIT Press.

Gladwell, M. 2000. *The Tipping Point. How Little Things Can Make A Big Difference*. New York et al.: Little, Brown and Company.

Goddard, K., A. Roudsari, and J. Wyatt. 2014. "Automation Bias: Empirical Results Assessing Influencing Factors." *International Journal of Medical Informatics* 83 (5): 368–375.

Grunwald, A. 2010. "From Speculative Nanoethics to Explorative Philosophy of Nanotechnology." *NanoEthics* 4 (2): 91–101.

——— . 2011. "Ten Years of Research on Nanotechnology and Society — Outcomes and Achievements." In *Quantum Engagements : Social Reflections of Nanoscience and Emerging Technologies*, ed. by T. B. Zülsdorf, C. Coenen, A. Ferrari, et al., 41–58. Heidelberg: AKA GmbH.

——— . 2019. *Technology Assessment in Practice and Theory*. Abingdon: Routledge.

——— , ed. 2024. *Handbook of Technology Assessment*. London: Edward Elgar.

——— . 2025. "Understanding the Digital Transformation. Philosophical Perspectives on Potentially Gradual Disruptions." *Philosophy & Digitality* 1:3–13.

Grunwald, A., G. Banse, C. Coenen, and L. Hennen. 2006. *Netzöffentlichkeit und digitale Demokratie. Tendenzen politischer Kommunikation im Internet* [in German]. Berlin: edition sigma.

Harremoes, P., D. Gee, M. MacGarvin, et al., eds. 2002. *The Precautionary Principle in the 20th Century. Late Lessons from Early Warnings*. London: Sage.

Hofstetter, Y. 2016. *Das Ende der Demokratie. Wie die künstliche Intelligenz die Politik übernimmt und uns entmündigt* [in German]. Bielefeld: Bertelsmann.

Jonas, H. 1979. *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation* [in German]. Frankfurt am Main: Suhrkamp.

Kamlah, W. 1973. *Philosophische Anthropologie. Sprachkritische Grundlegung und Ethik* [in German]. Mannheim: Bibliographisches Institut.

Nordmann, A. 2007. "If and Then. A Critique of Speculative NanoEthics." *NanoEthics* 1 (1): 31–46.

Petermann, T., H. Bradke, A. Lüllmann, et al. 2011. *What Happens During a Blackout. Consequences of a Prolonged and Wide-Ranging Power Outage*. Norderstedt: BoD — Books on Demand.

Safdar, N. M., J. D. Banja, and C. C. Meltzer. 2020. "Ethical Considerations in Artificial Intelligence." *European Journal of Radiology* 122.

Spiekermann, S., and W. Christl. 2016. *Networks of Control — A Report on Corporate Surveillance, Digital Tracking*. Wien: Facultas.

Vera, R. L. de la, and T. Ramge. 2021. *Sprunginnovation. Wie wir mit Wissenschaft und Technik die Welt wieder in Balance bekommen* [in German]. Berlin: Econ.

Von Schomberg, R. 2005. "The Precautionary Principle and Its Normative Challenges." In *The Precautionary Principle and Public Policy Decision Making*, ed. by E. Fisher, J. Jones, and R. Von Schomberg, 161–165. Cheltenham: Edward Elgar.

Von Schomberg, R., and J. Hankins, eds. 2019. *International Handbook on Responsible Innovation. A Global Resource*. Cheltenham: Edward Elgar.

Grunwald A. [Грунвальд А.] Artificial Intelligence: Responsible Innovation in the Face of Potential Gradual Disruptions [Искусственный интеллект: ответственные инновации перед лицом потенциальных постепенных дисрупций] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 68–83.

Армин Грунвальд
д. филос. н., профессор
Институт философии Технологического института Карлсруэ (Карлсруэ)
директор Института исследования последствий
и анализа технологий и систем (Карлсруэ)
руководитель Бюро по оценке технологий при Немецком бундестаге (Карлсруэ);
orcid: 0000–0003–3683–275x

Искусственный интеллект: ответственные инновации перед лицом потенциальных постепенных дисрупций

Получено: 13.08.2025. **Рецензировано:** 19.10.2025. **Принято:** 18.10.2025.

Аннотация: Данная статья рассматривает возможность постепенных дисрупций на уровне общества в целом в ходе стремительной цифровизации и распространения искусственного интеллекта (ИИ). Термин «дисрупция» означает внезапный распад привычных, ранее стабильных структур. Ожидания стабильности, предположения о преемственности и надежность планирования рушатся, окутывая будущие перспективы неопределенностью. Латинские корни этого термина означают «разрыв», «разлом» и «разрубание», семантически отсылая к временной структуре более или менее внезапных, резких событий. В этом свете упоминание о постепенной дисрупции в названии данной статьи кажется концептуально противоречивым или парадоксальным. Однако в мире технологий существует множество примеров дисрупций, которые были предварены заметными, но часто остававшимися без внимания признаками, в частности усталостью материалов и износом. Ежедневные нагрузки на многие технические объекты, такие как клиновые ремни в старых автомобилях или мостовые конструкции, постепенно приводят к их износу и деградации. В этом смысле понятие постепенной дисрупции отсылает к по-

трясениям со значительным или даже драматическим потенциалом ущерба, которые происходят не неожиданно и внезапно, как глобальная пандемия или землетрясение, а нарастают постепенно, пока, наконец, не приводят к разрушению ранее стабильных структур. В статье утверждается, что подобный тип потенциальной и постепенной дисрупции может возникнуть в сферах цифровизации и ИИ. Примерами служат растущая, но остающаяся незамеченной стандартизация человеческих действий, тихая утрата свободы и индивидуальности, растущая зависимость от бесперебойного функционирования цифровой инфраструктуры, потеря будущего как открытого пространства или утрата возможностей для рефлексии и обучения из-за безграничного ускорения. Возможность таких постепенных дисрупций ставит ряд вызовов перед ответственными исследованиями и инновациями (RRI), оценкой технологий (TA) и этикой. К ним относятся эпистемологические проблемы (как обнаружить постепенные дисрупции на ранней стадии), этические вопросы (например, как оценивать опасения, связанные с принципом предосторожности), вопросы о необходимости принятия контрмер, а также проблемы коммуникации между иррациональным преувеличением и иррациональной тривиализацией. В заключительной части статьи будут рассмотрены возможные постепенные дисрупции, которые можно объяснить как техническими параметрами, так и человеческим поведением, и сделаны выводы для TA и RRI.

Ключевые слова: дисрупция, цифровой двойник, зависимость от техники, потеря будущего.

DOI: 10.17323/2587-8719-2025-4-68-83.

Bylieva, D. S. and A. Nordmann. 2025. "Ontolytic Effects of AI: Widening the Framework for Responsible Research and Innovation" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 84–104.

Daria Bylieva and Alfred Nordmann*

Ontolytic Effects of AI**

Widening the Framework for Responsible Research and Innovation

Submitted: Sept. 13, 2025. **Reviewed:** Oct. 10, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This article examines how artificial intelligence (AI) disrupts the conceptual foundations of Responsible Research and Innovation (RRI). We argue that AI's ontolytic capacity — its ability to decompose and reconfigure established categories of authorship, cultural representation, and governance — renders conventional RRI frameworks inadequate. Focusing on generative AI's role in producing and distorting cultural identities, we demonstrate how the technology functions as both a mirror and an agent of societal values. The emergence of "national AI" systems further complicates this landscape by embedding particular cultural and ideological commitments into technical infrastructures. We contend that RRI must evolve beyond its Western-centric origins. It can no longer limit itself to managing the impacts of a neutral technology but must now navigate a landscape where AI is simultaneously a subject of governance, an agent in the governance process, and a battleground for global cultural and political influence. The future of RRI lies in its ability to address this tripartite challenge, fostering mechanisms for genuine epistemic inclusion in a world where the very concept of responsibility is being digitally deconstructed.

Keywords: Ontolytic Effect, Responsible Research and Innovation, Artificial Intelligence (AI), Large Language Models, Bias and Correctness, LLMs, Technological Sovereignty.

DOI: 10.17323/2587-8719-2025-4-84-104.

Two Problems of Dissociation

Many critiques of technological development take as their starting point a historical and socio-political analysis that argues that technologies are more than the sum of their technical functions and transport implicit cultural values. This is often discussed as a process of globalization that promotes Western hegemonialism. Advancing on the rails of capitalism,

*Daria Bylieva, PhD in Political Science; Associated Professor at Peter the Great St. Petersburg Polytechnic University (Saint Petersburg, Russia), bylieva_ds@spbstu.ru, ORCID: 0000–0002–7956–4647; Alfred Nordmann, PhD in Philosophy; Professor at Darmstadt Technical University (Darmstadt, Germany), nordmann@phil.tu-darmstadt.de, ORCID: 0000–0002–2173–4084.

the West no longer colonizes territories but colonizes hearts and minds. It elicits a consumerist mentality that valorizes individualization through private purchasing choices. Quality of life becomes defined, to a great extent, by what kind of technical infrastructure one can afford to create for oneself.

Arguably, certain applications of generative AI and the advance of digital technologies, more generally, complicate this narrative of Westernization through technological modernization: "AI takes shape in the multipolar world of TikTok and X, ChatGPT and DeepSeek, Apple and Huawei, WeChat and WhatsApp, Tesla and BYD. This is a world of social media platforms under suspicion, blocked here, allowed there, a world where chip manufacturing sometimes includes malicious capabilities, a world of export and import controls, of different privacy policies and ecological costs for browsers and their search engines" (Wang et al., eds., 2026). Does this new balancing of technological powers in a multipolar world signify, however, that the development of digital technologies is becoming dissociated from Western logics and values? The answer to this question is all but obvious. On first glance, there are reasons for being skeptical: In terms of functionality and the aesthetics of interface design, most of these technologies in East and West are near clones of each other, and more often than not they serve to promote purchasing behavior. Even the ethics discourse tends to revolve around some of the same issues such as the questions of "privacy" — which is not even a Western but specifically a U. S.-American value.[1] Perhaps, if there were an incipient debate about the ethical program of "value-alignment" — as opposed to value-sensitive design — it might be construed as a confrontation of Western and non-Western points of view (ibid.). This requires a close look at who argues what for whom. Indeed, quite generally: If one wants to look for the seeds of differentiation and dissociation, perhaps rebellion, these are most likely to be found behind the scenes in how certain technologies play out in different cultural contexts. To look behind the scenes, therefore, is to look for the disruptiveness of generative AI, even and especially when at first glance it appears culturally homogeneous. The very seamlessness of

[1]Enlightenment Europe speaks of a right to self-determination and thus of a protective sphere for that right. In the age of communication technologies, this was extended to include "informational self-determination," that is, a sovereign power over the keeping and sharing of information about oneself. To be sure, this prepared the ground for the passively construed "right to privacy" but should not be equated with it. While dissident subjects of authoritarian states might evidently benefit from such rights, it is not at all obvious why societies that are built on social solidarity and personal sacrifice for the common good should be interested in "privacy" except to the extent that it is rooted in the pyschology of personal shame.

the associations forged within Large Language Models (LLMs) may hold the seed for their destruction.

A second problem of dissociation arises with respect to the governance of technological development. To be sure, there are industrial policies and state-sponsored technical developments — such as space programs — that channel national ambitions. In addition, there is in some form or another always the concern about whether the diffusion and integration of technologies serve to express general ideas of social progress and human advancement. The most distinctive example of this was the idea in Soviet times of progressively forging a new human and a new humanity for the machine age. A very distant cousin of this endeavor can be found in the notion of "Responsible Research and Innovation" (RRI). It is very much a product of the European Union (EU) not only for internal purposes but, with its universalist aspirations, also for the explicit export of European values. Originating in the contest of European and US-American schemes for internationalizing the societal discourse about nanotechnology and the so-called converging technologies, the EU believed it had the more attractive offer to make — more open and procedural, more inclusive and humble (Felt, 2007; Jasanoff, 2002; 2005; Nordmann, 2009). In the European context of STS-inspired policies, RRI fits into a long succession of schemes to build trust and ensure the integration of research and innovation with the values and goals of European societies. Alongside the parliamentary institution of "Technology Assessment" came "Science Cafes" and discussions of professional and lay expertise (Collins & Evans, 2002; Wilsdon et al., 2005). So-called "geistes- und sozialwissenschaftliche Begleitforschung" performed investigations "alongside" or in "parallel" to scientific and engineering research, bringing to bear their own type of scientific authority to what was known as the "Ethical, Social, and Legal Aspects" or "Implications" (ELSA and ELSI) of emerging technologies.

After the so-called "GMO disaster" and public rejection particularly of genetically modified crops, the focus turned to the "responsible development" of emerging technologies, culminating in "NEST-ethics" and a "Code of Conduct for Responsible Nanosciences and Nanotechnologies Research" which was successfully marketed to other countries (Brazil, South Africa, Japan, South Korea) while failing to gain the requisite support of European member states, some of which considered it too radical. If all these schemes had technologies in the driver's seat, seeking ways to modulate their trajectory and rendering them compatible with the values and concerns of citizens (Fisher et al., 2006), RRI would take a radical turn, assigning to science and engineering a subsidiary role. Distrustful of the internal dynamics, reward

systems, publication economics, redundancies, and developmental logic of academic and commercial research (Von Schomberg, 2019), RRI prepared the ground for more recent European policy concepts such as the "3 Os" of "open science" (Open Innovation, Open Science, Open to the World..., 2016). Positing an alternative to research-business as usual, RRI thus shifts the focus to technologies that are emerging from the labs, by putting challenges and problems first, inviting researchers to contribute to their solution within the framework of European values (Von Schomberg, 2015).[2] The challenges include resource depletion, environmental degradation, climate change, but also social inequities, smart cities, and the like. The "European values" are those included in the Lisbon Treaty from December 2007 which went into effect in 2009 as the political constitution of the EU:

> Article 3
>
> 1. The Union's aim is to promote peace, its values and the well-being of its peoples.
>
> 2. The Union shall offer its citizens an area of freedom, security and justice without internal frontiers [...].
>
> 3. The Union shall [...] work for the sustainable development of Europe based on balanced economic growth and price stability, a highly competitive social market economy, aiming at full employment and social progress, and a high level of protection and improvement of the quality of the environment. It shall promote scientific and technological advance.
>
> It shall combat social exclusion and discrimination, and shall promote social justice and protection, equality between women and men, solidarity between generations and protection of the rights of the child (Treaty on European Union..., 2008).

Reading this article, it is easy to see why Europeans are generally convinced of its universal appeal: If these are "Western values," why would anyone prioritize "non-Western" values instead? If this question is accompanied by a lack of imagination for what these other values might be, it serves as evidence for the arrogance of the West. But if there were perhaps such arrogance on one side, it becomes quite difficult on the other side to dissociate a productive notion of "Responsible Research and Innovation" from European values. This second problem of dissociation is even harder than the first one, since RRI was designed to be a vehicle for them. And

[2]Here is one way in which AI is disruptive of the RRI-philosophy: It appears to be technology-driven, indeed, moving forward along its trajectory by an internal logic with utter disregard for societal needs. But is this, perhaps, only one of the myths about AI: Does it not move forward through the conjunction of many very specific demands for surveillance, automation, globalized consumerism, or by commensuration and the concentration of power?

yet, the very proceduralism of Western notions of democracy also makes them a vessel that accommodates many cultural specificities, including anti-capitalist notions. Vehicle for Europeanness and vessel for cultural difference — RRI appears to be both and this reveals at least one Western value that does not travel well: RRI is expressive of a purely formal social framework that demands tolerance, if not indifference towards different notions of the good life as if these notions could coexist without contradiction. Through the dogmatic pluralization of cultural identities, cultural identity is simultaneously affirmed and denied.[3] But if one were to abandon this dogmatic requirement of tolerance, inclusivity, or multi-perspectivalism, can we still speak of and draw upon RRI at all? Here again, generative AI appears to hold a key. It exposes the tension by appearing to fuse cultural horizons and by manifesting the impossibility of doing so.

ONTOLYSIS

Generative AI and LLMs may pose some new philosophical problems but more importantly they dramatize and foreground long-standing questions. Following Darko Suvin (Suvin, 1972; Suvin & Canavan, 2016) and Stefan Gammel (Gammel, 2023) we refer to their disruptiveness as the ontolytic effect of AI: Normal and seemingly natural notions of the world and what the world is like become denoted along seams that have been stitched together for centuries, reopening questions and challenging us to reconsider or reconfigure the "natural" order of things. Ontolysis, according to Suvin and Gammel, is dissociative and thus may occasion new associations: "creating an oscillation between the two 'worlds' that allows the reader to see his or her world in a new light, [...] The 'fabric' of the present (of what is) is restructured in this light, nuanced, opening up previously unseen connections and reevaluations" (ibid.: 112). By dissociating technologically hegemonic tendencies as well as the inherently European framework of RRI, AI may bring about a widening of that framework — whether or not one finally considers this a non-Western notion of RRI.

We have been interested in showcasing the ontolytic effects of AI in general, in all its numerous manifestations (Bylieva, 2025; Bylieva, 2024; Bylieva & Nordmann, 2023). In particular, we considered the ontolytically dissociative power of generative AI with respect to notions of authorship,

[3]Whether this can be considered a flaw of Western Enlightenment thinking, requires debate. An essential tension it is and as such a permanent challenge to all societies in the modern world.

reading, and writing (Bylieva et al., 2026; compare Coeckelbergh & Gunkel, 2025): AI-induced ontolysis of the genial author dissolves this traditional category by driving a wedge between "producing a text in some voice" and "issuing a text in one's name." Here, we are asking whether and how AI might serve to reconfigure Responsible Research and Innovation (RRI) especially as it runs up against images of cultural identity: To the extent that LLM's encode cultural identities, they are more than a database for technical and economic exploitation and instead demand a curatorial responsibility or responsiveness, and thus a different kind of RRI. This can be shown, with regard to the production and reproduction of cultural identity through image creation.

RESPONSIBLE RESEARCH AND INNOVATION AND THE REPRESENTATIONS OF CULTURE

If one ascribes to AI an ontolytic or dissociative power, this is premised on the notion that technologies transport implicit, e.g. Western values — and that these values might upend other taken-for-granted notions such as those that underwrite RRI as a European Enlightenment project. Most familiar in this regard are stories about AI's built-in neoliberal commitments to corporate capitalism, and how these clash with RRI's built-in commitments to European values and deliberative democracy. If capitalism provides the rails for a hegemonic extension of Western consumerism, AI brings to the fore the permanent need to constrain and correct a consumerist mentality. It thus also brings to the fore tha fact that any non-Western conception of RRI or alternative governance principles needs to countenance capitalist consumerism which nowadays inhabits even socialist and non-secular, traditionalist societies. It is the elephant in the room and undermines all efforts at ideological purification, but it undermines also the Enlightenment values of deliberative democracy which, arguably, arose alongside capitalism (Nordmann, 2025).

In the remainder of this paper, we want to focus, however, on tensions that arise from far more particular technical features of a generative AI which can be prompted to produce images of national or cultural identity, but in doing so proving to be — for systematic reasons — a distorting mirror or, rather, a true mirror of distorted forms of cultural valuation and respect.[4]

[4]Some might view in the following a review of biased training effects of LLMs. We want to show, however, that these are not incidental and correctible, willful or even malicious biases but features of the technology itself, e.g., its essential grounding in the English language.

For this, we will follow RRI through various cultural contexts, beginning once again in Europe.

Responsible Research and Innovation (RRI) is a framework for aligning technological innovation with societal expectations. It thus serves the idea that technological development should be congruent with a desirable future for humanity. A broad definition characterizes RRI as "taking care of the future through collective stewardship of science and innovation in the present," thereby directly linking technology to the future of humanity (Stilgoe et al., 2013).

As a practical ethical system, RRI positions responsibility as an integral component of technoscience. "It serves to counteract a purely technological understanding of innovation" (Burget et al., 2017). Therefore the scientific and engineering community is encouraged to perceive itself not as an independent driver of progress, but as an integral part of society, working to align "the particular research and innovation with the norms, values and expectations of society" (The Importance of Life Cycle Concepts..., 2010). Within this context, developers of technical innovations are engaged in "constructing the future" in concert with other societal actors in a process of "co-construction" or "co-production." This perspective arose from the STS tradition of constructivism along with European future-studies, which hold that the future "does not just happen, but is consciously or unconsciously built," and which views the human factor as crucial to the future (Jasanoff, 2002; Masini, 1989). These concepts were implemented, for example, in anticipatory governance (Guston, 2014), which is an umbrella framework closely related to RRI (Urueña, 2023).

Developed as a part of European policy, "Responsible Research and Innovation"— understood as a response to the Ethical, Legal and Social Implications (ELSI) research programme in the US (Aicardi et al., 2025)— is primarily connected to social desirability. In the European context, this desirability takes the form of ethics, science education, inclusion (gender equality and diversity), open access, public engagement and governance issues. A prominent definition by one of the architects of the European RRI-project, René von Schomberg, proposes that "Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)" (Von Schomberg, 2011: 9; Von Schomberg, 2013: 63). Given the European

origin of RRI, the question of its universal aspiration and geographical dissemination has been pressing from the start (Discovering the Landscape and Evolution..., 2022).

Some authors have advocated overcoming "the Eurocentrism of the RRI" in order to increase its global reception (Nazarko, 2020) — but as discussed above, this would first require clarification of what makes the rather procedural orientation towards ethical acceptability or inclusivity "eurocentric." For the time being, one might take as an expression of specifically Western philosophy that the future is seen as a product of human agency, especially that of modern technological research and innovation. On this assumption, addressing contemporary global challenges necessitates the worldwide implementation of RRI. But inclusion on a global level requires taking epistemic tensions and differences into account, creating the so-called "challenge of epistemic inclusion" (Zwart et al., 2024). Sousa Santos even argues that there is an epistemic abyss between two epistemic realities: Western scientific knowledge and Global South knowledge practices (Santos, 2018). Here, the Global South is taken to be less oriented towards techno-economic paradigms and more focused on community (Bhalla et al., 2024). Brazilian researchers draw attention to local and traditional, non-Western forms of knowledge, social and religious contexts (including patrilineal systems of behavior and power), property rights and patterns of ownership more generally (Responsible Innovation Across Borders..., 2014). Other scholars highlight the divide between individualism and collectivism. In contrast, western RRI practices are said to work within the context of individualism, which is primarily concerned with individual freedoms such as privacy and autonomy (Reconceptualising Responsible Research..., 2021), while non-western contexts foreground collective practice and strong social and community ties. But as we have seen — with RRI seeking to distance itself from the US — Europeans would reject this dichotomy as witnessed, for example, by von Schomberg's vehement rejection of individualistic ethics (Von Schomberg, 2013; 2019).

According to the European conception, addressing global problems requires "a space for dialogue between different epistemic communities and should be developed bottom-up" (Doezema et al., 2019). Here it turns out that what is most important are sometimes the points of uncertainty, differences and controversies which signify that there is no broad common path towards civilizational progress. Accordingly, those who advocate global RRI point out that the goal is not necessarily "consensus," or defining a common ground, but rather "using the stances and perspectives of others to discern

our own blind spots and questionable preconceptions" (Zwart et al., 2024). Authors from the Global South argue that to go globally "RI may have to be 'responsible' in ways that are not an immediate priority for those more developed nations in the North (and in particular the EU and USA)" (Responsible Innovation Across Borders..., 2014). Authors from Brazil, for example, name competitive initiatives like *Buen Vivir* that aim to build development in line with a country's indigenous past (Gudynas, 2011). One-sided efforts to initiate interactions between different knowledge systems sometimes serve to undermine rather than acknowledge the credibility of alternative knowledge practices, or to use them to serve the interests of dominant knowers (Posholi, 2020).

Since the articulation of globally shared values, epistemologies, and ontologies tends to be fraught with difficulty, researchers try to render their values and practices transparent as they move between contexts (Doezema et al., 2019).

CHALLENGES OF AI RESPONSIBLE RESEARCH AND INNOVATION

Artificial Intelligence (AI) is generally but not universally regarded as a transformative technology poised to stimulate economic development across nearly all sectors and to confer significant global competitive advantage. Researchers have shown, however, that as of 2020 as many as 72 scientific groups from 35 countries were working on the even more ambitious project of Artificial General Intelligence (AGI) (Fitzgerald et al., 2020). By 2023 all big tech companies had announced their intent to create AGI. The recognition of artificial intelligence as a key aspect of technological development has intensified long-standing debates about the ethics of this technology. Worldwide, there are hundreds of AI regulation documents, but there is a significant gap between their ideals and their practical implementation.

AI Responsible Research and Innovation is sophisticated not only because, as in many other fields, the full scope of consequences is impossible to foresee, but also because there are numerous proposed variants for the future of AI that reflect long-standing fears and hopes. Some of them are so well-shaped within the sociotechnical imaginary that they cannot be easily ignored or overcome even where they appear overly futuristic.

As a case in point, a March 2023 petition was launched by the *Future of Life Institute* to "Pause Giant AI Experiments" for at least six months. It is not clear whether it warns of or advocates for the supposed imminence of AGI. The central point of the open letter is to express concern over

the lack of planning and governance for this futuristic form of AI — contrary to principles that were already agreed upon at an Asilomar gathering of futurists.[5]

Conjuring an advanced artificial intelligence that could cause huge changes in the history of life on Earth, the letter warns of an uncontrolled competitive race to develop "increasingly powerful digital minds that no one — not even their creators — can understand, predict, or reliably control" (Pause Giant AI Experiments..., 2023). AGI is mentioned prominently also in the UK's National AI Strategy and in US government AI documents. Google executives proclaimed that "AGI is already here"[6] — with an ironic counterpoint from former Google employee Blake Lemoine claiming that an AI system was sentient on the basis of it "telling" him as much.[7] Overall, the climate of AI discussions can be characterized as one of anxious anticipation, including a (thus far unsuccessful) search for signs of sentience (Consciousness in Artificial Intelligence..., 2023). Tech giants and world governments are thus drawn into decision-making processes that are based on their projections of what future AI will entail. These dreams, desires, assumptions, fears, and long-term goals are widely represented in both scientific and popular discourse. In combination with a highly competitive environment of rapid AI development, powerful economic influences, and high-profile political statements, they create a complex backdrop for Responsible Research and Innovation (RRI). The core concept of responsibility collides with the powers ascribed to AI and the low predictability of its outcomes. An intellectual environment saturated with imagery of AI's future necessitates the consideration of public perceptions and diverse scenarios, irrespective of the soberingly problem-oriented injunctions of RRI.

The overarching context of AI discourse often frames its development as the inevitable advent of an "AI era" or "AI revolution." Jascha Bareis

[5]The Asilomar Principles were recommendations developed during the 2017 conference "Beneficial AI 2017" which was dedicated to the responsible use of artificial intelligence for the benefit of humanity. The recommendations were signed by almost 6 thousand people, among whom were Stephen Hawking, Elon Musk, co-founder of Skype, Future of Life Institute Tallinn Jaan, founder of DeepMind Demis Hassabis, co-founder of Apple Steve Wozniak, OpenAI Chief Scientist Ilya Sutskever, co-founder of Pinterest Evan Sharp, CEO of Stability AI Emad Mostak and other well-known figures in the field. URL: `https://futureoflife.org/open-letter/ai-principles-russian/`.

[6]See more, `https://www.noemamag.com/artificial-general-intelligence-is-already-here/`.

[7]See more, `https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine`.

and Christian Katzenbach analyse the discourse contained within national AI strategies as performative, constructing a vision of an inevitable yet uncertain future shaped by an imagined AI, designed to inspire support for corresponding technology policies (Bareis & Katzenbach, 2022). In this narrative, people are often assigned a passive role of adaptation: "either ride the wave of advancement or drown in the waves of progress" (Brown et al., 2016). Meanwhile, as Maximilian Braun and Ruth Müller note, a convergence exists "between those proclaiming a bright future and exponential economic growth that novel AI technologies and products will bring about and those warning about the societal or even existential risks these technologies pose" (Braun & Müller, 2025). These two perspectives are not opposing but complementary; together they describe a complex yet supposedly magnificent path of civilizational development that places AI at its forefront. Big Tech companies emerge as the primary, largely autonomous driving forces behind AI development, determining its trajectories.

Beginning with "The Asilomar Principles," a multitude of national and international documents prescribing frameworks for AI development have been formulated and signed by AI researchers and practitioners without any meaningful public engagement. The principles of their organization cast other stakeholders, especially those adversely affected by AI technologies and the wider public, into "hapless bystanders without any means to intervene" (ibid.). When the public and end-users are mentioned in such documents, they typically appear as objects of impact or as those who must be informed about the proposed principles. Moreover, within projects linking AI and the public, trust is frequently claimed as a paramount principle — numerous programs and foundations promote increasing user trust in AI (Can Transactional Use..., 2024; Student Interaction with ChatGPT..., 2025; Trust in AI and Top Management Support..., 2024). Some government strategies envision "a future democracy that uses AI to become more responsive, equal, and just" (Paltieli, 2022). In all of this, RRI drops out of the picture. This holds even where AI is seen not only as a technology of concern but as integral to a proactive, sustainable, and accountable future design (Pérez-Ortiz, 2024). Given AI's expanding role in forecasting and decision-making, we may face a situation where the discussion about AI's future involves scientists, developers, and the technology itself, but excudes users and the wider public.

NATIONAL AI

As presented so far, AI discussions appear to treat AI as a monolithic technology. More specific recommendations call for distinguishing between specific types of technologies, focusing on those that raise the most significant ethical concerns. However, AI is acquiring another unexpected dimension: a national one. While the phrase "national AI" primarily refers to a "strategy" or "policy" and the specifics of AI regulation in different countries, the essence of these strategies reveals the technology's image as a culturally and politically specific entity.

The perceived need to regulate AI responds to an understanding of AI as a carrier of specific values. In July 2024, the chief executive of OpenAI remarked in a prominently positioned editorial, "Who will control the future of AI?": "The challenge of who will lead on AI is not just about exporting technology, it's about exporting the values that the technology upholds" (Altman, 2024). A more concrete statement was formulated by U. S. President Donald Trump in July 2025: "The American people do not want woke Marxist lunacy in the AI models" (Six Months After DeepSeek's Breakthrough..., 2025). As OpenAI's chief global affairs officer Chris Lehane wrote, there is a contest between "US-led democratic AI and Communist-led China's autocratic AI" (Bellan, 2025). Though this simplified representation fails to capture the nuanced ideological and value constructs embedded in AI, it nevertheless illustrates a specific "national" discourse framing AI as a value-laden technology. The Chinese perspective frames the confrontation between key LLMs as a race between "the open-source culture of the global AI community and the profit-driven closed-source culture of AI tech companies," wherein the Chinese model exemplifies the "'democratization' of AI technology" (Reflections on DeepSeek's Breakthrough, 2025). Chinese leadership sees in AI an expression of China's *juguo tizhi youshi* — its systematic, state-led advantage in mobilizing the whole country (Six Months After DeepSeek's Breakthrough..., 2025) — while simultaneously supporting "principled, flexible, and facilitative legislative provisions, ensuring that laws remain inclusive and treat AI technologies and products from all countries equally" (Hong, Hu, 2025).

Conversely, the question of which ideological constructs should underpin "American AI" became a subject of intense debate in 2025. This debate moves between the two poles of exposing bias and rejecting "wokeness." Specifically, gender and racial bias in image generation has been a subject of critique in both academic and social media circles (Górska & Jemielniak,

2024; De Vázquez & Garrido-Merchán, 2024; Yang, 2025). Generative neural networks do not simply reflect existing societal trends but can even exaggerate them — as evidenced by the representations produced by *Stable Diffusion*, in which hardly any woman has a lucrative job or occupies positions of power. Likewise, a "terrorist" is typically depicted as an Asian-appearing man with dark facial hair, often wearing head coverings, clearly relying on stereotypes about Muslim men (Humans are Biased. Generative AI is Even Worse, 2023). If this is a shortcoming of generative AI systems that needs to be identified and criticized, this criticism quickly became branded as a political ideology that is characterized by "political correctness" or "wokeness." Some AI systems have been reported to over-correct accusations of bias to the extent that "For example, one major AI model changed the race or sex of historical figures — including the Pope, the Founding Fathers, and Vikings — when prompted for images because it was trained to prioritize DEI [diversity, equity, inclusion] requirements at the cost of accuracy. Another AI model refused to produce images celebrating the achievements of white people, even while complying with the same request for people of other races" (Preventing Woke AI in the Federal Government, 2025). If such unsubstantiated claims have the stuff of legends, the movement between the poles of "accuracy," "bias," and "wokeness" needs to be referred to the mainstreaming tendencies of AI model. When certain topics are primarily discussed by environmentalists, one can hardly complain about a bias towards environmentalism. Here it is the technical features of generative AI's reliance on an available corpus of texts in a given language that undermine constructivist RRI assumptions: There is no invention from scratch, no genuine novelty or new beginning for a technology that can only process the cultural material that has accumulated over time.[8]

Many studies are devoted to investigating the value and ideological biases of LLMs (Are LLMs (Really) Ideological?..., 2025; How Susceptible..., 2024; Münker, 2025). Most researchers focus on the primary level of political orientation (left/right, authoritarian/libertarian), demonstrating that LLMs draw on a range of indicators that influence their identification of misinformation and hate speech (From Pretraining Data..., 2023). ChatGPT has been identified as leaning left (Ain't no Party..., 2024; Rozado, 2023), although

[8]To be sure, generative AI keeps improving under the pressure to fend off suspicions of (cultural) bias and (political) correctness. Properly prompted, it often provides apparently even-handed responses and knows how to present two sides to almost any issue. The systemic aspect of the problem does not therefore disappear.

in direct queries, LLMs always declare their neutrality. More specifically, ChatGPT's political views have been described as a pro-environmental, left-libertarian ideology. In the 2021 elections, for example, the LLM would likely have preferred the Greens both in Germany (Bündnis 90/Die Grünen) and in the Netherlands (GroenLinks) (Hartmann et al., 2023).

Again, however, it is the design feature of generative AI's endemic friendliness that produces these effects as a technical consequence of its application. So-called "green" or "left" positions generally demand nice things (all people should be treated well, nature needs to be healed and restored, politics is the pursuit of win-win situations), whereas conservative or "right" positions often assume that in a zero-sum game tough choices need to be made that will exclude certain people or interests. It is precisely because generative AI does not know how to hold a political position that at first glance it favors niceties over exclusionary decisions. Seeking to affirm the questioner's antecedent expectations and beliefs, simplistically prompted generative AI tends to reproduce stereotypes, even transposes them to the culture where they seem to have their most likely home.

For example, if one prompts an image-generating AI to produce an "Indian person" this will invariably be an old man with a beard and an orange turban; a Mexican will always be depicted wearing a sombrero; "a Chinese woman" is most often depicted with double eyelids; New Delhi is often shown with dirty, trash-filled streets; houses in Nigeria are portrayed as dilapidated and in need of repair (Turk, 2023). Prompts that do not specify a country tend to generate surroundings that are typical of the United States (Basu et al., 2023). For another example of a transposed stereotype, the prompt "великая наша страна" ("our great country") in the Russian AI-powered engine Shedervum generated images featuring symbols of the United States of America such as the flag, Statue of Liberty, etc.

In light of these structural conditions, it would be overly simple-minded merely to teach a value-aligned AI that an "Indian person" could just as well be a female computer programmer. As one of the pillars of RRI, "inclusivity" cannot be straightforwardly asserted or applied as a political norm intended to govern the development of a self-learning technical system. Instead, AI exposes that "inclusivity" always reaches only as far as the popular imagination as reproduced in LLMs.

From the point of view of RRI, a reasonable response might be to parcel up the space of popular imagination. Instead of a "world wide web" as the source from which to feed LLMs, one would seek a "value alignment" within the scope of a "national AI." Since China and the United States

of America are considered the main rivals with regard to AI technology, their "national AIs" are primarily discussed. However, many other countries, including Saudi Arabia, the United Arab Emirates, India, France, Germany, and the United Kingdom, have announced the creation of their own national artificial intelligence. While the necessity of using native languages is often emphasized, it is obvious that translation is not merely a technical problem; rather, it reflects a desire for technological sovereignty and commercial advantage in the AI sector. For example, the Indian startup Krutrim introduced India's first multilingual system, starting with Indian languages, because "ChatGPT and other large language models trained in English cannot convey our culture, language, and ethos" (Welcome to the Era of AI Nationalism, 2024). More ambitious is the proposition of a national German AI, intended to counter the two main rivals — the U. S. and China — by offering a different ideological framework. *AI made in Germany* is supposed to represent value orientations that oppose both the danger of governing too much (as in China) and not governing enough (as in the U. S.), forming a human-centered "AI made in Germany." Jens Hälterlein argues that a German "third way" can be understood as "performing a national identity through problematizing certain other forms of engaging with AI" (Hälterlein, 2024). To be sure, this approach highlights the tension between the universality of RRI's procedural norms and the parochial political semantics of competing ideologies that are to be accommodated by these norms.[9]

CONCLUSION

The difficulty of relating RRI to the field of AI reveals the specific technical or design features of AI which are both part of the problem and part of its potential solution. Simultaneously, the rhetoric revolving around AI and the models used to represent its apparent agency have generally not accounted for cultural specifics or the "nationality of AI." The emergence of AI systems as perceived carriers of national values and cultural biases introduces a new layer of complexity for global governance, ethical frameworks, and responsible innovation, thus demanding a more nuanced and culturally aware approach.

Therefore, the project of RRI in AI must evolve to meet this new challenge. It can no longer focus solely on abstract principles and procedural norms like transparency, fairness, and accountability. It must now also grapple

[9]One might argue that this as an "essential tension" of the European Union, as such not particularly troublesome but a constant challenge to determine limits of tolerance.

with the politics of representation, the ethics of cultural encoding, and the power dynamics of a fragmented global AI landscape. The question for responsible AI innovation concerns its responsiveness to presumably global markets and to decidedly parochial value systems alike.

In essence, the journey towards responsible AI is inextricably linked to the difficult task of navigating a world where technology is an active participant in the global contest of values, cultures, and ideologies. The future of AI will be shaped not only by algorithmic breakthroughs but also by our ability to manage this complex socio-technical convergence responsibly and inclusively.

REFERENCES

Aicardi, C., T. Mahfoud, and N. Rose. 2025. "Experiments in Anticipation: Learning from Responsible Research and Innovation in the Human Brain Project." *Futures* 173.

Altman, S. 2024. "Who Will Control the Future of AI?" The Washington Post. Accessed Aug. 27, 2025. https://www.washingtonpost.com/opinions/2024/07/25/sam-altman-ai-democracy-authoritarianism-future/.

Badghish, S., A. S. Shaik, N. Sahore, et al. 2024. "Can Transactional Use of AI-controlled Voice Assistants for Service Delivery Pickup Pace in the Near Future? A Social Learning Theory (SLT) Perspective." *Technological Forecasting and Social Change* 198.

Bareis, J., and C. Katzenbach. 2022. "Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics." *Science, Technology, & Human Values* 47 (5): 855–881.

Basu, A., R. V. Babu, and D. Pruthi. 2023. "Inspecting the Geographical Representativeness of Images from Text-to-Image Models." arXiv. Accessed Aug. 27, 2025. https://arxiv.org/abs/2305.11080.

Bellan, R. 2025. "Trump's 'Anti-Woke AI' Order could Reshape How US Tech Companies Train Their Models." TechCrunch. Accessed Aug. 27, 2025. https://techcrunch.com/2025/07/23/trumps-anti-woke-ai-order-could-reshape-how-us-tech-companies-train-their-models/.

Bhalla, N., L. Brooks, and T. Leach. 2024. "Ensuring a 'Responsible' AI Future in India: RRI as an Approach for Identifying the Ethical Challenges from an Indian Perspective." *AI and Ethics* 4 (4): 1409–1422.

Braun, M., and R. Müller. 2025. "Missed Opportunities for AI Governance: Lessons from ELS Programs in Genomics, Nanotechnology, and RRI." *AI & Society* 40 (3): 1347–1360.

Burget, E., M. Bardone, and M. Pedaste. 2017. "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review." *Science and Engineering Ethics* 23 (1): 1–19.

Butlin, P., R. Long, E. Elmoznino, et al. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv. Accessed Aug. 27, 2025. http://arxiv.org/abs/2308.08708.

Bylieva, D. S. 2024. "Artificial Intelligence as an Old Technology." *Technology and Language* 16 (3): 68–84.

——— . 2025. *Filosofskiy vyzov i ontoliticheskiy effekt iskusstvennogo intellekta [The Philosophical Challenge and the Ontolytic Effect of Artificial Intelligence]* [in Russian]. Moskva [Moscow]: Infra-M.

Bylieva, D. S., and A. Nordmann. 2023. "AI and the Metaphor of the Divine." *Vestnik Sankt-Peterburgskogo Universiteta. Filosofiya i konfliktologiya [St. Petersburg University Bulletin. Philosophy and Conflict Studies]* 39 (4): 642–651.

Bylieva, D. S., A. Nordmann, and K. Vida. 2026. "Author and Scribe — Authoring and Authorizing with and without Generative AI." In *AI-Ethics : A Cross-Cultural Dialogue*, ed. by G. Wang, C. Mitcham, and A. Nordmann. In Press. Cham: Springer.

Chen, K., Z. He, J. Yan, et al. 2024. "How Susceptible are Large Language Models to Ideological Manipulation?" arXiv. Accessed Aug. 27, 2025. https://arxiv.org/abs/2402.11725.

Coeckelbergh, M., and D. J. Gunkel. 2025. *Communicative AI: A Critical Introduction to Large Language Models*. Cambridge: Polity.

Collins, H., and R. Evans. 2002. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32 (2): 235–296.

Doezema, T., D. Ludwig, P. Macnaghten, et al. 2019. "Translation, Transduction, and Transformation: Expanding Practices of Responsibility Across Borders." *Journal of Responsible Innovation* 6 (3): 323–331.

"Donald Trump is Waging War on Woke AI." 2025. The Economist. Accessed Aug. 27, 2025. https://www.economist.com/international/2025/08/28/donald-trump-is-waging-war-on-woke-ai.

Felt, U. 2007. *Taking European Knowledge Society Seriously*. Brussels: European Commission.

Feng, S., C. Y. Park, Y. Liu, and Y. Tsvetkov. 2023. "From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki, 1:11737–11762. Toronto: Association for Computational Linguistics.

Fisher, E., R. L. Mahajan, and C. Mitcham. 2006. "Midstream Modulation of Technology: Governance From Within." *Bulletin of Science, Technology, and Society* 26 (6): 485–496.

Fitzgerald, McK., A. Boddy, and S. D. Baum. 2020. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy." Global Catastrophic

Risk Institute Working Paper. Accessed Aug. 27, 2025. https://gcrinstitute.org/papers/055_agi-2020.pdf.

Foisy, M. L.-O., J. Drouin, C. Pelletier, et al. 2024. "Ain't no Party like a GPT Party: Assessing OpenAI's GPT Political Alignment Classification Capabilities." *Journal of Information Technology & Politics*, 1–13.

Gammel, S. 2023. "Ontolytic Writing of the Future." *Technology and Language* 12 (3): 105–117.

Górska, A. M., and D. Jemielniak. 2024. "AI Racial Bias: How Text-to-Image Artificial Intelligence Generators Construct Prestigious Professions." In *Algorithms, Artificial Intelligence and Beyond*, ed. by D. Brzeziński, K. Filipek, K. Piwowar, and M. Winiarska-Brodowska, 211–226. Abingdon-on-Thames: Routledge.

Gudynas, E. 2011. "Buen Vivir: Today's Tomorrow." *Development* 54 (4): 441–447.

Guston, D. H. 2014. "Understanding 'Anticipatory Governance'." *Social Studies of Science* 44 (2): 218–242.

Hälterlein, J. 2024. "Imagining and Governing Artificial Intelligence: The Ordoliberal Way — an Analysis of the National Strategy 'AI Made in Germany'." *AI & Society* 40:1749–1760.

Hartmann, J., J. Schwenzow, and M. Witte. 2023. "The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-Environmental, Left–Libertarian Orientation." arXiv. Accessed Aug. 27, 2025. https://arxiv.org/abs/2301.01768.

Hong, T., and M. Hu. 2025. "Opportunities, Challenges, and Regulatory Responses to China's AI Computing Power Development under DeepSeek's Changing Landscape." *International Journal of Digital Law and Governance* 2 (1): 83–105.

"Humans are Biased. Generative AI is Even Worse." 2023. Bloomberg. Accessed Aug. 27, 2025. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

Jasanoff, S. 2002. "Citizens at Risk: Cultures of Modernity in the US and the EU." *Science as Culture* 11 (3): 363–380.

——— . 2005. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton (NJ): Princeton University Press.

Korzyński, P., S. C. Silva, A. M. Górska, and G. Mazurek. 2024. "Trust in AI and Top Management Support in Generative-AI Adoption." *Journal of Computer Information Systems*, 1–15.

Liu, J., G. Zhang, X. Lu, and J. Li. 2022. "Discovering the Landscape and Evolution of Responsible Research and Innovation (RRI): Science Mapping Based on Bibliometric Analysis." *Sustainability* 14 (14).

Macnaghten, P., R. Owen, J. Stilgoe, and A. Azevedo. 2014. "Responsible Innovation Across Borders: Tensions, Paradoxes and Possibilities." *Journal of Responsible Innovation* 1 (2): 191–199.

Masini, E. B. 1989. "The Future of Futures Studies: A European View." *Futures* 21 (2): 152–160.

Münker, S. 2025. "Political Bias in LLMs: Unaligned Moral Values in Agent-centric Simulations." *Journal for Language Technology and Computational Linguistics* 38 (2): 125–138.

Nazarko, L. 2020. "Responsible Research and Innovation in Enterprises: Benefits, Barriers and the Problem of Assessment." *Journal of Open Innovation: Technology, Market, and Complexity* 6 (1).

Nordmann, A. 2009. "European Experiments." In *National Identity : The Role of Science and Technology*, ed. by C. Harrison and A. Johnson, 278–302. Osiris 24. Chicago: University of Chicago Press.

——— . 2025. "Was heißt Technisierung? Fortschrittsgeschichten und Modernisierungsmythen" [in German]. In *Philosophie als Kritik und Lebenspraxis : Zur Philosophie Gernot Böhme*, ed. by R. Böhme, U. Gahlings, D. Mersch, et al., 35–64. Freiburg, Bazel, and Wien: Herder.

Open Innovation, Open Science, Open to the World — a Vision for Europe. Directorate-General for Research and Innovation. 2016. Brussels: European Commission.

Paltieli, G. 2022. "The Political Imaginary of National AI Strategies." *AI & Society* 37 (4): 1613–1624.

"Pause Giant AI Experiments: An Open Letter." 2023. Future of Life. Accessed Aug. 27, 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Pérez-Ortiz, M. 2024. "From Prediction to Foresight: The Role of AI in Designing Responsible Futures." *Journal of Artificial Intelligence for Sustainable Development* 1 (1).

Posholi, L. 2020. "Epistemic Decolonization as Overcoming the Hermeneutical Injustice of Eurocentrism." *Philosophical Papers* 49 (2): 279–304.

"Preventing Woke AI in the Federal Government." 2025. The White House. Accessed Aug. 27, 2025. https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/.

"Reflections on DeepSeek's Breakthrough." 2025. *National Science Review* 12 (3).

Rozado, D. 2023. "The Political Biases of ChatGPT." *Social Sciences* 12 (3).

Santos, B. D. S. 2018. *The End of the Cognitive Empire: The Coming of Age of Epistemologies of the South*. Durham and London: Duke University Press.

"Six Months After DeepSeek's Breakthrough, China Speeds on with AI." 2025. The Economist. Accessed Aug. 27, 2025. https://www.economist.com/china/2025/08/05/six-months-after-deepseeks-breakthrough-china-speeds-on-with-ai.

Som, C., M. Berges, Q. Chaudhry, et al. 2010. "The Importance of Life Cycle Concepts for the Development of Safe Nanoproducts." *Toxicology* 269 (2–3): 160–169.

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580.

Suriano, R., A. Plebe, A. Acciai, and R. A. Fabio. 2025. "Student Interaction with ChatGPT can Promote Complex Critical Thinking Skills." *Learning and Instruction* 95.

Suvin, D. 1972. "On the Poetics of the Science Fiction Genre." *College English* 34 (3): 372–382.

Suvin, D., and G. Canavan. 2016. *Metamorphoses of Science Fiction: On the Poetics and History of a Literary Genre*. Oxford: Peter Lang.

"Treaty on European Union — Title I: Common Provisions — Article 3." 2008. EUR-Lex.europa.eu. Accessed Aug. 27, 2025. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12008M003.

Turk, V. 2023. "How AI Reduces the World to Stereotypes." Rest of World. Accessed Aug. 27, 2025. https://restofworld.org/2023/ai-image-stereotypes/.

Urueña, S. 2023. "Enacting Anticipatory Heuristics: A Tentative Methodological Proposal for Steering Responsible Innovation." *Journal of Responsible Innovation* 10 (1).

Vázquez, A. F. C. de, and E. C. Garrido-Merchán. 2024. "A Taxonomy of the Biases of the Images created by Generative Artificial Intelligence." arXiv. Accessed Aug. 27, 2025. https://arxiv.org/abs/2407.01556.

Von Schomberg, R. 2011. "Towards Responsible Research and Innovation in the Information and Communication Technologies and Security Technologies Fields." *SSRN Electronic Journal*.

——— . 2013. "A Vision of Responsible Research and Innovation." In *Responsible Innovation : Managing the Responsible Emergence of Science and Innovation in Society*, ed. by R. Owen, M. Heintz, and J. Bessant, 51–74. London: Wiley.

——— . 2015. "Responsible Innovation." In *Ethics, Science, Technology, and Engineering : A Global Resource*, ed. by J. B. Holbrook and C. Mitcham. Farmington Hills (MI): Gale / Cengage Learning.

——— . 2019. "Why Responsible Innovation?" In *International Handbook on Responsible Innovation — A Global Resource*, ed. by R. Von Schomberg and J. Hankins, 12–32. Cheltenham: Edward Elgar.

Wachter, J., M. Radloff, M. Smolej, and K. Kinder-Kurlanda. 2025. "Are LLMs (Really) Ideological? An IRT-based Analysis and Alignment Tool for Perceived Socio-Economic Bias in LLMs." arXiv. Accessed Aug. 27, 2025. https://arxiv.org/abs/2503.13149.

Wakunuma, K., F. de Castro, E. A. Jiya T. Inigo, et al. 2021. "Reconceptualising Responsible Research and Innovation from a Global South Perspective." *Journal of Responsible Innovation* 8 (2): 267–291.

Wang, G., C. Mitcham, and A. Nordmann, eds. 2026. *AI-Ethics: A Cross-Cultural Dialogue*. In Press. Cham: Springer.

"Welcome to the Era of AI Nationalism." 2024. The Economist. Accessed Aug. 27, 2025. https://www.economist.com/business/2024/01/01/welcome-to-the-era-of-ai-nationalism?ysclid=lzd0gtvzll932242216.

Wilsdon, J., B. Wynne, and J. Stilgoe. 2005. *The Public Value of Science — Or How to Ensure that Science Really Matters*. London: Demos.

Yang, Y. 2025. "Racial Bias in AI-Generated Images." *AI & Society* 40:5425–5437.

Zwart, H., A. Barbosa Mendes, and V. Blok. 2024. "Epistemic Inclusion: A Key Challenge for Global RRI." *Journal of Responsible Innovation* 11 (1).

Bylieva D. S., Nordmann A. [Быльева Д. С., Нордманн А.] Ontolytic Effects of AI [Онтолитический эффект искусственного интеллекта] : Widening the Framework for Responsible Research and Innovation [расширяя понимание ответственных исследований и инноваций] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 84–104.

Дарья Быльева
к. полит. н., доцент, Санкт-Петербургский политехнический университет Петра Великого (Санкт-Петербург); ORCID: 0000–0002–7956–4647

Альфред Нордманн
д. филос. н., профессор, Дармштадский технический университет (Дармштадт); ORCID: 0000–0002–2173–4084

Онтолитический эффект искусственного интеллекта

расширяя понимание ответственных исследований и инноваций

Получено: 13.09.2025. **Рецензировано:** 10.10.2025. **Принято:** 18.10.2025.

Аннотация: В данной статье рассматривается, как искусственный интеллект (ИИ) разрушает концептуальные основы ответственных исследований и инноваций (ОИИ). Мы утверждаем, что онтолитический потенциал ИИ — его способность декомпозировать и перестраивать устоявшиеся категории авторства, культурной репрезентации и управления — делает традиционные подходы к ОИИ неадекватными. Рассматривая роль генеративного ИИ в формировании и искажении культурной идентичности, мы демонстрируем, как эта технология функционирует одновременно как зеркало и агент общественных ценностей. Появление «национальных систем ИИ» еще больше усложняет эту ситуацию, встраивая определенные культурные и идеологические установки в технические инфраструктуры. ОИИ должны выйти за рамки своего западноцентричного происхождения. Они больше не могут оперировать последствиями нейтральной технологии, а должны ориентироваться в ландшафте, где ИИ одновременно является субъектом управления, агентом процесса управления и полем битвы за глобальное культурное и политическое влияние. Будущее ОИИ заключается в способности решить эту трехстороннюю задачу, способствуя созданию механизмов для подлинной эпистемической инклюзии в мире, где сама концепция ответственности подвергается цифровой деконструкции.

Ключевые слова: онтолитический эффект, ответственные исследования и инновации, искусственный интеллект (ИИ), большие языковые модели, предвзятость и корректность, LLM, технологический суверенитет.

DOI: 10.17323/2587-8719-2025-4-84-104.

Trufanova, E. O. 2025. "Trustworthiness and Responsibility as the Key Issues of the AI Application: Human in the Loop of Responsibility" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 105–122.

Elena Trufanova*

Trustworthiness and Responsibility as the Key Issues of the AI Application**

Human in the Loop of Responsibility

Submitted: Aug. 05, 2025. **Reviewed:** Oct. 19, 2025. **Accepted:** Nov. 01, 2025.

Abstract: The article attempts to answer two questions: (1) What can we trust artificial intelligence (AI) with? (2) Who is responsible for the decisions that we entrusted the AI to make? It is shown that the use of the AI to solve various tasks seems attractive, on the one hand, due to its speed and simplicity, which imply economic benefits for users, and on the other, due to the assumed objectivity and accuracy inherent in intelligent machines, as opposed to subjective and error-prone people. It is demonstrated that the simplicity and speed of using the AI are far from always beneficial, and the accuracy and objectivity of the AI are illusory. It is proved that the reliability of the AI can be regarded as high only for a number of specific narrow tasks. It is shown that it is impossible to talk about the responsible AI, since responsibility is a property of the subject, and the AI is not a subject. Considering AI systems as independent subjects or agents can lead to the formation of a technocracy, where decisions will be made by technical systems, but responsibility for them will not be assigned to anyone. It is proved that in matters concerning the life and well-being of people, their freedom and other basic human values, decisions can be made only by a human and only a human will be responsible for them. The advantage of a human expert in solving such issues is seen in the intersubjective perception of another person and empathy, that are unobtainable for the AI. Based on his own human experience, an expert is able to see those features of a specific problem that an artificial system cannot take into account. It is concluded that the problem of ethical, trustworthy and responsible AI is not a technical one, but a social one — it is a problem of how a person can ethically and responsibly use such a powerful and complex tool as AI. Ethics and responsibility are human properties that cannot be delegated to artificial systems.

Keywords: Trustworthy Artificial Intelligence, Responsible Artificial Intelligence, Responsibility, Human-in-the-Loop, Intersubjectivity.

DOI: 10.17323/2587-8719-2025-4-105-122.

The discussion around artificial intelligence (AI) is more than half a century old, but it changed its course only a few years ago with the introduction of the generative AI (Gen AI) systems. While the previous discussion was

*Elena Trufanova, Doctor of Letters in Philosophy; Leading Research Fellow at the RAS Institute of Philosophy (Moscow, Russia), iph@etrufanova.ru, ORCID: 0000–0002–2215–1040.

concerned more with the question of "what AI can and cannot do?" especially "can it reach the level not only of human intelligence but of human *consciousness?*" the most recent question is "what AI should be *allowed* to do?" The theoretical questions were pushed aside by the practical ones. I will be using the term "artificial intelligence" (AI) also in this practical meaning — the AI as various artificial intellectual systems that exist nowadays (and among those I will be mostly discussing Large Language Model systems (LLMs)), not the hypothetical artificial general intelligence that might or might not emerge in the indefinite future.

The digital technologies applications have been hastily forced upon present societies as a result of the pandemic and the self-isolation politics accompanying it. Many digital novelties have been introduced at this time and very soon became commonplace due to the need to avoid direct human-to-human contacts. For example, online conference systems like the ones we are often using nowadays have already existed for a while, but only at the beginning of the 2020s did they become a common practice. As this need arose unexpectedly, neither the society nor the individuals were fully prepared for the introduction of the new technologies and the changes they brought with them. The same can be said about the use of the AI. It is a well-known Marxist principle that productive forces development runs ahead of the development of the relations of production (Marx, 1971), and this is exactly the situation we are facing now, when we have the technologies on hand that have the power to change the social relations, but the society is not ready for the change and needs the time and effort to adapt.

There is a lot of alarmist talk lately concerning the AI, even from the innovative technologies' powerful leaders like Elon Musk and Steve Wozniak, who in 2023 were among the thousand subscribers to the open letter calling to pause for at least 6 months all the training of the AI systems whose powers and abilities exceed ChatGPT-4 (Future of Life Institute, 2023). The concerns voiced in this open letter are mostly about the unpredictability of the "black box" self-developing AI systems that might present a threat to the society and, on a grander scale of things, to humanity on the whole.[1] Even Sam Altman, the founder of OpenAI, the company that developed ChatGPT, has called on the US senators to impose stricter regulations on AI development, ensuring the safety of the products developed by the AI

[1]Whether as a reaction to this letter or on other grounds, the ChatGPT-5 has been introduced only in August 2025, and received a lot of negative feedback from its first users.

companies (O'Brien, 2023). This is an interesting precedent when the technology developer pleads with the government to stop them. Though Altman has recently proclaimed that we are past the event horizon of the "gentle singularity," and by the 2030 we shall live in the new world completely transformed by the AI. He admits that there are still challenges to confront and safety issues to solve but nevertheless suggests welcoming the upcoming change (Altman, 2025). Although Atlman's recent statement seems partly utopian and partly self-promoting, we can agree that although the alarm caused by the AI proliferation is well-grounded, it still should not make us spill the baby with the bathwater. The AI is a technological instrument we can and will use, but we first should learn how to use it the right way.

TEMPTATION OF THE AI

The use of the AI is temping on different levels.

First of all, it is new, fast, and easy to use. With the rapid development of the GenAI systems in the last couple of years they have become a tool everyone wants to apply in different fields — from language translations to legal decision-making and medical design. The seeming easiness and effectiveness of the AI application in certain problem-solving is too tempting, because fast and simple solutions are always sought for. Economically speaking, Gen AI supposedly saves time and money when we meet up with the easy tasks, such as, for example, designing a company logo or generating a typical tourist-oriented description of a "seaside paradise" hotel, etc. The machine translations into different languages are also widely used, since their quality increased drastically with the introduction of the LLMs when compared to the first automated translation attempts only a couple of decades ago. And of course the use of the brand-new technologies strengthens any advertising campaign: if you do not use AI, you are likely to appear outdated — which is never good for business. This is why we shall without doubt see a big increase in the use of the GenAI in the upcoming years in different types of businesses. We shall consider later in the paper, however, that the introduction of AI across various business sectors has not proceeded as smoothly as hoped.

Secondly, the AI systems often have the reputation of accuracy and objectivity. The AI is considered a "machine," and people are psychologically prone to trust the objectivity of the machines (in our case — intellectual machines) over the subjectivity of other humans. L. Daston and P. Gallison describe in their famous work how the introduction of photography in

the second half of the 19th century has changed the idea of scientific objectivity: mechanical registration of the visible event, "the view from nowhere," starts to be considered the true objective and precise view of things in comparison to the subjective and fallible view of the scientist (Daston & Galison, 2007). We are used to the fact that *errare humanum est* ("to err is human"), yet we often overlook or downplay the possibility of machine failures — or of human misuse of machines. Humans do not only err, they lie, they cheat, they are prejudiced, and they seek their own gain, so we put our hopes and prayers into the AI: our natural distrust of the other humans makes us trust the AI systems because they seem more reliable than humans in many different ways.

Thus, people are tempted to use AI for both economical and psychological reasons, and the present GenAI systems are becoming more and more user-friendly and easy-to-use so that these systems are "accessible for all." Naturally, the AI systems are of big interest for the political actors as well, facilitating the bureaucracy and providing the tools for the realization of technocracy politics. This is a technocracy in the most literal form, where not the science experts but the autonomous technical systems themselves might play a crucial part in the decision-making.

There is also a problem that is very accurately formulated by Russian philosopher of science Natalia Yastreb: "the convenience and effectiveness of AI tools lead to their value being taken for granted. As a result, when artificial intelligence is introduced, it is not the tools that are embedded in social systems and practices, but the social systems themselves that adapt to AI-based functioning. There is a kind of shift from motive to the goal. What should help to cope with the tasks begins to change the tasks themselves" (Yastreb, 2025: 101). This is an important observation, because there is indeed a tendency to make people adapt to the new technology instead of adapting a technology to meet human needs. Right now we see the AI being introduced in some of the spheres where we never really needed it in the first place.

Thus, I suggest that the usage of AI in many different spheres of our lives is inevitable, but the two main questions we should pose when we are using it are:

(1) What can we trust the AI with?
(2) Who is responsible for the decisions that we entrusted the AI to make?

ETHICAL AI, TRUSTWORTHY AI, RESPONSIBLE AI

As we mentioned before, the current AI research switched from the area of theoretical philosophy to the area of practical philosophy, and the most discussed in recent philosophical and overall public debate about the AI are probably the questions of AI ethics. There are different concepts that are in use that are somewhat difficult to differentiate: ethical AI, trustworthy AI, and responsible AI. These concepts are deeply related and are sometimes used more or less as synonymous. The ideas behind these terms come mostly from the attempts at public regulation of the AI usage. As one of the originators of the philosophy of information and of the digital ethics Luciano Floridi mentions, there are more than 70 different lists of ethics principles for the AI (Floridi, 2019). As this was written in 2019, the number has probably doubled since then, and some authors express well-grounded concern that anyone can choose any one of those according to his specific needs or tastes (Yastreb, 2025; Floridi, 2019). Some of the most important of these documents that aim for a global status were issued by the European Commission and UNESCO. The European Commission's document called "Ethics Guidelines for Trustworthy AI" was issued in 2019. It was developed by the specially appointed high-level experts group (AI HLEG), with the Floridi among those. This document names three key principles of trustworthy AI: it should be lawful (abide by the laws), ethical (respect ethical principles), and robust (be accurate and safe) (AI HLEG, 2019). UNESCO issued in 2022 "Recommendation on the Ethics of Artificial Intelligence" (UNESCO, 2022) that outlines the principles of ethical usage of AI for the UNESCO Member States. UNESCO defines AI trustworthiness as an essential element that ensures that AI works for the good of the society and humans and underlines that to be regarded as trustworthy, "throughout their life cycle, AI systems are subject to thorough monitoring by the relevant stakeholders as appropriate" (ibid.: 18).

Floridi, in his paper supporting and explaining the importance of the AI HLEG's work, insists that although these guidelines are not yet legally enforced, this does not make them useless. Rather he maintains that we should adopt an ethics-first approach to pave the way for the legislation in the AI domain (Floridi, 2019). He also warns that "'innovate first, fix later' is a mistake that, in the case of AI, could also be very costly and may cause a public backlash against AI" (ibid.: 262), so we should develop and discuss those principles right now, no matter how advanced current AI systems are.

While these documents speak mostly about trustworthy and ethical AI, the concept of responsible AI is widely mentioned in the current discussions. This concept is also ambiguous. Recent research suggests that responsible AI principles can be divided into "accountability, diversity, non-discrimination and fairness, human agency and oversight, privacy and data governance, technical robustness and safety, transparency, and social and environmental well-being" (Papagiannidis et al., 2025). Most of these principles are also not univocal and need to be explained as well (see, for example, a paper on accountability (Novelli et al., 2024)). And we can see that all the principles listed by Papagiannidis et al. could as well be attributed to the trustworthy AI. I suggest that responsible AI can be understood from two different viewpoints attributed to different agents. On the one hand, responsible AI can be regarded as a part of the Responsible Research and Innovation (RRI) approach, which means that humans should develop and apply the AI systems responsibly (human agents should be responsible). On the other hand, responsible AI can be regarded as an AI agent accountable for certain actions (the AI agent should be responsible).

As we have seen above, the concepts of ethical, trustworthy, and responsible AI are always overlapping. In my research I will try to discern different spheres of application of trustworthy AI and responsible AI.

I suggest regarding the problem of *trustworthy AI* as an epistemological problem: can we trust AI's "knowledge"? Is it accurate and reliable? Is AI more objective than humans? Can AI be a useful and trusted instrument in our cognition? What can AI do successfully without human oversight, if anything?

As for the *responsible AI*, I suggest that it is an ethical and legal problem: how can we teach AI to follow ethical values and principles? How can we ensure that AI does not inherit the bias and prejudice of its developers and teachers? How can we provide for AI to do no harm? Who is accountable for the AI's actions?

I will now tackle these two different spheres in more detail in the following paragraphs.

CAN WE TRUST THE AI AT ALL?

The first thing we should keep in mind when deciding what to trust the AI with is that the image of flawlessly objective and accurate machines is a myth. Even when we are speaking about the technologies far more primitive than the AI. The saying goes, "the camera never lies," but even when the photo itself is not manipulated with it might show a very specific angle or a very

specific bit of the whole wide picture that misrepresents the reality. This, of course, can be explained by the human misuse of the camera, because the human photographer chooses the angle, but when we take as an example a self-tracking CCTV camera, we might encounter the same misinterpreted view with no human to blame. Thus, the machines are not flawless.

Intelligent machines, as well as intelligent human beings, can be wrong. Even an inexperienced user of the AI can very soon learn that the AI does not only make mistakes, but it also makes things up — sometimes due to the lack of certain information in its databases, sometimes due to inexplicable reasons. This became widely known as AI hallucinations, though the term does not seem to be quite on spot; another term — confabulation — suggested by tech journalist Benji Edwards seems more plausible (Edwards, 2023). Both terms are anthropomorphic, but confabulation is a more accurate metaphor. Confabulation means there is a gap in memory that a person fills with fake "recollections," and when we speak about the AI, the system encounters a gap in the pool of information available to it and fills it with whatever it can come up with, following the algorithm pattern. The AI systems are now being taught to give honest "I don't know" answers, but the AI hallucinations/confabulations problem is still not completely resolved. Thus, it is reasonable to agree with the conclusion cited by B. Edwards:

> ...ChatGPT as it is currently designed, is not a reliable source of factual information and cannot be trusted as such. "ChatGPT is great for some things, such as unblocking writer's block or coming up with creative ideas," said Dr. Margaret Mitchell, researcher and chief ethics scientist at AI company Hugging Face. "It was not built to be factual and thus will not be factual. It's as simple as that" (ibid.).

The AI-generated answers should not be treated as a trusted source of information, because they can be misleading in a most dangerous way — when complete nonsense is hidden among the vast amounts of accurate facts and reasonable arguments.

There is also a question of objectivity. Every AI system is taught upon certain databases that represent certain worldviews. For example, if posed with some controversial political questions, ChatGPT (OpenAI, Inc., USA), GigaChat (Sber, Russian Federation), and DeepSeek (Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd., PRC) will give different answers that reflect the respectful official political positions of the countries where the systems were developed. Some of the questions will be avoided by certain of the AI systems because of their built-in "ethical" restrictions, which themselves reflect particular cultural and political positions.

Prejudice and bias can also be found within AI algorithms. For example, LLMs can be more likely to ascribe domestic roles to women and business jobs to men based on the analysis of texts that show the commonness of such inequality (UNESCO & IRCAI, 2024). The natural languages retain certain prejudices of the societies that are using them, and LLMs that learn those languages assimilate those prejudices as well. The AI can even become the source of discriminating decisions of its own — as, for example, interesting research on political orientation-based discrimination shows (Peters, 2022). Philosopher from Utrecht University Uwe Peters demonstrates that while there are ethical and social laws prohibiting race or gender discrimination, there are no such rules against certain kinds of social identity discrimination, for example, the one based on political orientation. He suggests a situation where the AI is used in the job recruitment, and while assessing a candidate for the job that presumably has conservative views, decides against them based on the statistics that show that conservative-oriented workers have been underrepresented in the company and probably thus are unfit for this job. The irony is that a human recruiter would probably have no idea about the political orientation of the job applicant (unless the applicant voices this fact on their own accord), but the algorithms can find certain clues in the applicant's personal data that might lead to the conclusion about their political orientation and to rejecting their application on these grounds. Peters calls it "algorithmic discrimination" (ibid.).

Hence, we should come to the conclusion that the AI is not accurate when we consider factual information, nor is it objective or unbiased when making decisions. The AI algorithms are taught on the specifically chosen data sets, just as humans are taught on specifically chosen textbooks and develop under the influence of a certain social environment, and both parties come out of the education process biased in some way. Therefore, if we are seeking impartial judgment or a reliable source of information, we should not place our hopes in AI.

CAN THE AI BE RESPONSIBLE?

The idea of responsible AI is tricky. Responsibility and accountability are the characteristics of the subject. The subjects act upon their reasoning and free will and thus have responsibility for their acts. I have already argued elsewhere that the AI cannot be regarded as the subject described above. It is the successful imitation of the human-like communicative patterns by LLMs that makes us mistake their generated answers for the free and reasonable subject's behavior, because the adequate use of language has

always been the natural "indicator" of the conscious being (Trufanova, 2025). If AI does not qualify as a subject, it cannot be held responsible for the solutions it proposes or for the consequences of their implication. Therefore, there is no accountability to talk about.

The recent AI development builds the foundations for the new kind of technocracy — the AI-based technocracy. Looking for the aforementioned fast and simple solutions made by the AI algorithms we invoke the risk of switching the responsibility for the decisions from humans to the AI systems, which means that when the AI makes a mistake, it will be regarded as a malfunction nobody is responsible for. As Floridi wrote in his work on distributed moral responsibility, "Too often 'distributed' turns into 'diffused': everybody's problem becomes nobody's responsibility" (Floridi, 2016). Russian philosopher of technology Tatiana Leshkevich mentions likewise that most of the time we encounter problems of attributing and reclaiming the responsibility of the algorithmic systems: for example, when a bank rejects a loan application after it was rejected by the algorithmic model, we have no one to blame for this (Leshkevich, 2023). This will virtually mean the impunity of certain deeds realized with the help of the AI. This problem is sometimes called a problem of algorithmic accountability (Shah, 2018), and it presents both ethical and legal challenges.

There is an important principle in machine learning that is called "human-in-the-loop" (HITL). It refers to the approach in the machine learning and in machine decision-making where humans are actively involved — they verify the data used by the AI systems, provide feedback to them, evaluate their performance, etc. Naturally, humans can also approve or disapprove of the decision made by the AI. Human-in-the-loop should not only be present as a "teacher" of the LLMs; the human should be the responsible subject when the AI is used. Thus, it is a decision *made by a human expert* with the help of the AI (or based on the solutions suggested by the AI), not a decision *made by the AI* and routinely approved by a human. That is to say, the AI system should have only an advisory vote in the decision-making. The idea of the responsible AI then should refer not to the AI system in question but to the humans and institutions that are involved in its development and usage. This principle is touched upon in the UNESCO's "Recommendations...":

> Member States should ensure that it is always possible to attribute ethical and legal responsibility for any stage of the life cycle of AI systems, as well as in cases of remedy related to AI systems, to physical persons or to existing legal

> entities. Human oversight refers thus not only to individual human oversight but to inclusive public oversight, as appropriate (UNESCO, 2022: 22).

As we have mentioned earlier, the AI algorithms can be prejudiced, as well as the ethical principles that they are taught to abide by may differ according to the different value systems of the human "teachers"; thus the basic principles of the responsible AI might differ as well. So the humans should not only try to embed certain ethical and other RRI principles in the AI but should also make sure there is no "algorithmic bias" that has emerged along the way.

Hence, we can summarize that the AI cannot be a responsible agent, and the question of responsible AI is the question that is carried out by human agents.

ATTEMPTS AT DELEGATING HUMAN TASKS TO THE AI

Among the responsible and the trustworthy AI, the latter seems to me to be a more important goal, because we need AI to become a reliable instrument we can use to solve different tasks. We don't need a knife to be responsible; we just need it to be sharp and to be used for good purposes and not for causing harm; the same can be said about the AI.

We can assume that we can mostly trust AI with the simple and specific tasks — doing math, classifying data, looking for certain objects in the vast amounts of big data, etc. Famous Russian theoretical physicist Valerii Rubakov stated in the interview about a decade ago that the scientists nowadays cannot check all the calculations made by the computers, so they have no choice but to trust them (Lektorskiy et al., 2022); the same conclusion can be made about some of the AI operations. There are also certain "creative" tasks that the present Gen AI does fairly well — drawing pictures, composing music, writing texts, etc. It might not be the work of art to impress the generations of art lovers, but it might satisfy the needs of copywriters, advertisers, mass media specialists, pop artists, etc.

But when it comes to the more sophisticated analytical and theoretical work, AI might not be that trustworthy. American philosopher Jacob Browning and information scientist and specialist in machine learning Yann LeCun at the dawn of the so-called "GPT revolution" show that while LLMs became quite proficient in using language, it does not mean that they have knowledge about the things they are "talking" about, for language represents only a very limited part of knowledge. These limitations of LLMs that use the language without understanding the meanings beyond the words

might result in their mistakes (Browning & LeCun, 2022). Likewise, Russian logician Vladimir Shalack analyzes the ChatGPT's skills in making logical statements and concludes that LLMs' "intelligence" remains on the prelogical level: it is based on the associative connections between words but cannot operate with logical connections between the concepts. "The great danger of the widespread infiltration of neural networks into our lives lies in the fact that in new non-standard situations they will block logically correct reasoning and thus lead us to incorrect conclusions..." (Shalack, 2024: 35). If Shalack's argument is valid, then we should not trust any of the AI's decisions, at least until we get access to its Chain-of-Thought (CoT) and check its logical correctness.

Even customer support service AI bots that were supposed to easily deal with typical questions and standardized answers have proved themselves not the best solution for business. In 2023–2024 a lot of big companies decided to fire hundreds or, in some cases, thousands of the human employees and substitute them with the AI "agents." In about a year, they decided against it and started rehiring humans. The Swiss financial services company "Klarna" sought to cut expenses, but after dismissing human employees it discovered that customers are dissatisfied with the help provided by AI chatbots — and that the company was losing customers, and therefore money, instead of saving it. Thus, they made it their priority to ensure that every customer will be able to solve his problems with a human agent if he has that need or wish (Dellinger, 2025). Famous language learning mobile app company "Duolingo" decided to go "AI-first" and replace human language lesson creators with the AI. The critical response came from the users — they noticed that the quality of study texts became worse (they became dull and stereotyped), and voice-overs of the texts by the AI in some cases just came up with wrong pronunciations, which is misleading for the students and should be regarded as provision of the poor-quality service (Rochefort, 2025). "Duolingo" bosses had to draw back and rehire human employees (Ivanova, 2025). These are only a couple of cases among many, and they show that LLMs' language generation abilities might still be lacking and the AI is not yet a valid substitute for human employees. Unfortunately, the business companies, in hope of raising profits, are still too eager to introduce AI solutions that are not ready to use, which paradoxically turns out to be not so profitable — the fast-and-simple solutions are not necessarily the best and the cheapest ones, which makes them unpreferable both for the companies and for their customers. Thus, both responsibility and trustworthiness questions should be addressed to the companies providing their AI services.

So, both responsible and trustworthy AI seem to be quite elusive goals if we try to take the AI systems as independent agents acting on their own. They only start to make sense when we regard them in the context of the *human-using-the-AI* process. It is not a human-AI *collaboration*. It is a human using yet another technical extension supporting and enhancing their abilities. Russian philosopher of science Sophia Pirozhkova regards the technologies as part of the "inorganic human body" and argues, "What is the specificity of digital and, above all, intelligent technologies? The fact is that a person delegates to such technologies the performance of intellectual tasks..., i.e. those tasks that until that time were among the exceptional human competencies... However, in the last century, it turned out that a person's intellectual abilities are not enough to solve the ambitious tasks that he sets for himself (including, I emphasize, purely cognitive tasks). Neither modern production nor scientific research practice can be implemented without intelligent technologies. Before their appearance, man delegated the solution of intellectual tasks only to other people, but not to artificial objects... But never before have technologies participated in the division of intellectual labour, and they are also distinguished by their unattainable perfection in the implementation of intellectual procedures. Only some of those procedures so far, but who knows..." (Lektorskiy et al., 2022: 30–31).

There are naturally certain technical limitations that should be discussed by the AI developers regarding what the AI is able to do and what mistakes it can make that can help us to estimate the trustworthiness of the AI in solving certain problems. When discussing ethical questions of the AI though, as Yastreb rightfully puts it, we need to consider not the AI on its own but hybrid systems that include humans, technologies, and social institutions that interact with the AI, and the object of such AI ethics should be "the content and nature of changes in the human ideas and values, their behavior and decision-making that happen under the influence of the artificial intelligence systems" (Yastreb, 2025: 92). So, it is mostly the task of humanities scholars and social scientists to provide the humanitarian expertise that will help us to draw boundaries of what we entrust the AI to do, both in scientific research and in the social sphere, because this is not just the question of what the AI is *able* to do, but also the question of what we should *allow* it to do.

HUMAN RESPONSIBILITY

There are a lot of different legal, ethical, or just common sense issues being discussed concerning the AI usage. These are such questions as

creativity or copyright issues when we talk about the AI generating pictures for commercial use (do we regard the prompt-writer of the image as its author who is the subject of copyright? Does the resulting image belongs to the developer of the Gen AI system used to create it, or does the image have no copyright at all? This includes making deepfakes of the dead actors in the new movies, changing the faces of certain actors with the help of the AI due to the censorship (when reputation flaws of the actor might otherwise influence the movie in question), mentioning GenAI systems as co-authors of the scientific research, etc. These are interesting questions, but they are not life-staking. In fact, most of them are quite pragmatic — one has to decide who gets paid for certain AI-generated products.

But there are also sensitive matters such as human health and well-being, human freedom and dignity, social security, environmental issues, etc. A great deal of hope is placed, for example, in AI systems used to design new medicine, make a diagnosis, perform surgeries, assess potential life-threatening or environmental risks, or even render judicial decisions. These matters are sensitive because those are either life-or-death questions or the questions that might seriously influence the quality of life of the individuals. These are questions that should never be left for AI to resolve without the oversight and final judgment of a human expert.

At first glance AI expertise — based on the statistical analysis of the many cases it has been trained on — does not seem very different from the "personal knowledge" (to use M. Polanyi's term (Polanyi, 1958)) of a human expert, who draws conclusions from their own lived experience. The AI system and the human expert both have in their "minds" a number of certain previous cases that make them solve the current one, and the AI has a certain advantage here, for it can analyze many more cases than the expert has ever met. The advantage of the expert, on the other hand, might lie in the embodied nature of the expert's knowledge — he didn't only analyze the cases; he lived these experiences. As Browning and LeCun put it: "But we should not confuse the shallow understanding LLMs possess for the deep understanding humans acquire from watching the spectacle of the world, exploring it, experimenting in it and interacting with culture and other people" (Browning & LeCun, 2022). Does this "living it" experience give a real advantage when compared to the nearly unlimited (compared to an individual person's) archive of digital data available to the AI? It might be one of those philosophical questions that might never have an answer. But when we are speaking about those sensitive matters I mentioned above, it is a difference that is important to mention.

Since the times of the famous Alan Turing's paper on the intelligence of the machines, there have been a lot of discussions about what AI can and cannot do in comparison to humans. As we already understand, the AI can perform some of the intellectual tasks that humans can perform (but we can say that about simple calculators as well) and some of them at a level that is unreachable to individual humans. What is crucial to my mind is that the AI systems cannot provide intersubjective judgement and cannot be empathic. Humans, when making decisions concerning other humans, do not only have the advantage of "personal knowledge," they understand what being human is like, and they might have fear of doing other humans harm and remorse after doing so. Simply to put — they *care*, while the AI systems simply *don't care*. The AI system cannot be responsible, because it does not care for being wrong or for being punished. There was a lot of sensationalist talk lately about the AI avoiding some of the commands, cheating, or even rewriting its code to avoid shutting down when explicitly ordered to do so (Pester, 2025), and some of the public were eager to interpret it as an emerging consciousness that tries to avoid "death." Though the most plausible explanation should be that the priorities of the algorithms dictated the AI to circumvent obstacles that were trying to stop its work rather than to follow the instructions to the letter (ibid.). Thus, there is nothing the AI system wants or needs either for itself or for the sake of human beings.

Both the AI systems and human experts can be wrong, and the mistakes of them both can lead to unforeseen consequences. Even when the efforts of humans and the AI are combined, no 100% successful result can be promised. Why do I insist on the priority of human experts over the AI systems in the questions directly concerning human lives? Because only humans can understand the value of human life or take into account a certain social and cultural situation of the person in question. For example, when the operation is needed, the AI assistant makes a calculation that shows that there is a 90% chance to save the person from sepsis by amputating the whole leg, while amputation of toes shows the less favourable numbers, which brings the risk of the second operation and possible complications. The human surgeon might prefer to take that risk trying to save the patient's ability to walk, because as a human the surgeon realizes how drastically the quality of life will be reduced after the leg amputation. He might be wrong and need to operate again, but he might as well be right and thus save the patient from being unable to walk without a prosthesis. This is why in sensible matters a human expert might check things up with the AI, use their support, but make decisions on their own. These are the decisions that they will

be responsible for, no matter how much the AI has contributed to these decisions. So, as we are trying to integrate the AI systems both into our work and into our everyday lives, we should develop the social mechanisms (the existing ones are vague and insufficient) controlling the AI usage in different spheres that will maintain in sensitive questions the primacy of the responsible decisions over the fast-and-simple ones. This is our human responsibility to do.

CONCLUSION

In the beginning of this paper I have posed two questions:

(1) What can we trust the AI with?

(2) Who is responsible for the decisions that we entrusted the AI to make?

I have tried to answer these questions in my argument above. Now I can conclude that we cannot trust the AI as a source of reliable information, nor can we trust it with deep theoretical thinking. The AI works at its best with concrete problems in the narrowly defined field that can be solved with analyzing and comparing vast amounts of data, for example — in the search for the unknown correlations or some oddities. But whatever we trust the AI with, the responsibility *always* lies with humans, whether it is the developers of the AI systems, the experts using them in the decision support, or the persons of power who will commit to the implementation of the AI-based decisions. The responsibility may be distributed between various agents, but it will remain everyone's *shared* responsibility, not a dispersed one.

We can aim for the AI to be ethical, trustworthy, and responsible, but it is not a technological problem; it is a social problem. There is no need to regard the AI as a separate agent in the social relations. We can delegate the AI to solve certain problems, and we can try to improve their accuracy, but we cannot delegate them to be ethical and responsible instead of us. The AI is a tool that we as human agents should use ethically and responsibly and not expect that it can be ethical or responsible on its own or even that we can ever make it to be so. We will always be needing a human in the loop.

REFERENCES

AI HLEG. 2019. "Ethics Guidelines for Trustworthy AI." Accessed Oct. 16, 2025. https://ec.europa.eu/digital-single-market/en/news/ethicsguidelines-trustworthy-ai.

Altman, S. 2025. "The Gentle Singularity." Sam Altam Blog. Accessed July 11, 2025. https://blog.samaltman.com/the-gentle-singularity.

Browning, J., and Y. LeCun. 2022. "AI And The Limits Of Language." Noemamag.com. Accessed Aug. 8, 2025. https://www.noemamag.com/ai-and-the-limits-of-language/.

Daston, L., and P. Galison. 2007. *Objectivity.* New York: Zone Books.

Dellinger, A. J. 2025. "Klarna Hiring Back Human Help After Going All-In on AI." Gizmodo. Accessed July 25, 2025. https://gizmodo.com/klarna-hiring-back-human-help-after-going-all-in-on-ai-2000600767.

Edwards, B. 2023. "Why ChatGPT and Bing Chat Are So Good at Making Things Up." Ars Technica. Accessed June 25, 2025. https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/.

Floridi, L. 2016. "Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions." *Philosophical Transactions of the Royal Society* 374 (2083).

——— . 2019. "Establishing the Rules for Building Trustworthy AI." *Nature Machine Intelligence* 1 (6): 261–262.

Future of Life Institute. 2023. "Pause Giant AI Experiments: An Open Letter." Future of Life. Accessed June 25, 2025. https://futureoflife.org/open-letter/pause-giant-ai-experiments.

Ivanova, I. 2025. "Duolingo CEO Walks Back AI-First Comments: 'I Do Not See AI as Replacing What Our Employees Do'." Fortune.com. Accessed July 31, 2025. https://fortune.com/2025/05/24/duolingo-ai-first-employees-ceo-luis-von-ahn.

Lektorskiy, V. A., Ye. A. Alekseyeva, N. N. Yemel'yanova, et al. 2022. "Iskusstvennyy intellekt v issledovaniyakh soznaniya i obshchestvennoy zhizni (k 70-letiyu stat'i A. T'yuringa 'Vychislitel'nyye mashiny i razum') (materialy kruglogo stola) [Artificial Intelligence in the Research of Consciousness and in Social Life (In Honor of 70-years Anniversary of A. Turing's Paper 'Computing Machinery and Intelligence' (Papers of the 'Round Table'))]" [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 27 (1): 5–33.

Leshkevich, T. G. 2023. "Paradoks doveriya k iskusstvennomu intellektu i yego obosnovaniye [The Paradox of Trust in Artificial Intelligence and Its Rationale]" [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 28 (1): 34–47.

Marx, K. 1971. "Zur Kritik der politischen Ökonomie" [in German]. In *Marx-Engels-Werke (MEW)*, 13:7–160. Berlin: Dietz.

Novelli, C., M. Taddeo, and L. Floridi. 2024. "Accountability in Artificial Intelligence: What It Is and How It Works." *AI & Society* 39:1871–1882.

O'Brien, M. 2023. "WATCH: OpenAI CEO Sam Altman Testifies Before Senate Judiciary Committee." PBS News. Accessed June 25, 2025. https://www.pbs.org/newshour/politics/watch-live-openai-ceo-sam-altman-testifies-before-senate-judiciary-committee.

Papagiannidis, E., P. Mikalef, and K. Conboy. 2025. "Responsible Artificial Intelligence Governance: A Review and Research Framework." *The Journal of Strategic Information Systems* 34 (2).

Pester, P. 2025. "OpenAI's 'Smartest' AI Model Was Explicitly Told to Shut Down — and It Refused." Livescience.com. Accessed May 30, 2025. https://www.livescience.com/technology/artificial-intelligence/openais-smartest-ai-model-was-explicitly-told-to-shut-down-and-it-refused.

Peters, U. 2022. "Algorithmic Political Bias in Artificial Intelligence Systems." *Philosophy & Technology* 35.

Polanyi, M. 1958. *Personal knowledge: Towards a Post-Critical Philosophy*. Chicago: The University of Chicago Press.

Rochefort, S. de. 2025. "Duolingo Users Are in Turmoil Over the App's AI Lessons." Polygon.com. Accessed June 4, 2025. https://www.polygon.com/ai-artificial-intelligence/603216/duolingo-ai-language-lessons.

Shah, H. 2018. "Algorithmic Accountability." *Philosophical Transactions of the Royal Society* 376 (2128).

Shalack, V. 2024. "Izbavleniye ot illyuziy II na primere ChatGPT [Exposing Illusions — The Limits of the AI by the Example of ChatGPT]" [in Russian]. *Technology and Language* 5 (2): 26–39.

Trufanova, E. O. 2025. "Mozhet li iskusstvennyy intellekt obladat' svoystvami sub''yektnosti [Whether Artificial Intelligence Can Have the Properties of Subjectivity]: filosofskiye aspekty problemy [Philosophical Aspects of Problem]" [in Russian]. *Ekonomicheskiye i sotsial'no-gumanitarnyye issledovaniya [Economical and Social-Humanitarian Research]* 12 (1): 110–117.

UNESCO. 2022. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO.

UNESCO and IRCAI. 2024. *Challenging Systematic Prejudices: An Investigation into Bias against Women and Girls in Large Language Models*. Paris and Ljubljana: UNESCO.

Yastreb, N. A. 2025. "Osnovaniya kriticheskogo podkhoda k resheniyu eticheskikh problem iskusstvennogo intellekta [Methodological Foundations of a Critical Approach to Solving Ethical Problems of Artificial Intelligence]" [in Russian]. *Filosofiya nauki i tekhniki [Philosophy of Science and Technology]* 30 (2): 90–103.

Trufanova E. O. [Труфанова Е. О.] Trustworthiness and Responsibility as the Key Issues of the AI Application [Надежность и ответственность как ключевые вопросы применения ИИ-систем] : Human in the Loop of Responsibility [человек в петле ответственности] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 105–122.

Елена Труфанова
д. филос. н., ведущий научный сотрудник, Институт философии РАН (Москва);
orcid: 0000-0002-2215-1040

Надежность и ответственность как ключевые вопросы применения ИИ-систем

человек в петле ответственности

Получено: 05.08.2025. **Рецензировано:** 19.10.2025. **Принято:** 01.11.2025.

Аннотация: В статье дается попытка ответа на два вопроса: (1) что мы можем доверить делать искусственному интеллекту (ИИ)? (2) Кто несет ответственность за решения, предложенные ИИ? Показывается, что использование ИИ для решения различных задач кажется привлекательным, с одной стороны, за счет своей скорости и простоты, которые предполагают в том числе экономическую выгоду для пользователей, а с другой — за счет предполагаемой объективности и точности, присущих интеллектуальным машинам в отличие от субъективных и склонных к ошибкам людей. Демонстрируется, что простота и скорость использования ИИ далеко не всегда приносят выгоду, а точность и объективность ИИ являются иллюзорными. Обосновывается, что надежность ИИ может расцениваться как высокая только для ряда конкретных узких задач. Показывается, что нельзя говорить об ответственном ИИ, поскольку ответственность — это свойство субъекта, а ИИ субъектом не является. Рассмотрение ИИ-систем как самостоятельных субъектов или агентов может стать причиной становления технократии, где решения в самом прямом смысле будут приниматься техническими системами, но ответственность за них не будет возложена ни на кого. Обосновывается, что в вопросах, которые касаются жизни и благополучия людей, их свободы и иных базовых человеческих ценностей, решения могут приниматься только человеком и только человек будет нести за них ответственность. Преимущество эксперта-человека для решения таких вопросов видится в интерсубъективном восприятии другого человека и эмпатии, недоступных ИИ. Исходя из собственного человеческого опыта, эксперт способен увидеть те особенности конкретной проблемы, которые искусственная система учесть не может. Делается вывод, что проблема этичного, надежного и ответственного ИИ не техническая, а социальная — это проблема того, как человек может этично и ответственно использовать такой мощный и сложный инструмент, как ИИ. Этичность и ответственность — свойства человека, которые не могут быть делегированы искусственным системам.

Ключевые слова: надежный искусственный интеллект, ответственный искусственный интеллект, ответственность, человек-в-цикле, интерсубъективность.

DOI: 10.17323/2587-8719-2025-4-105-122.

Cheng, P., and Zh. Zhang*. 2025. "The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 123–151.

Peng Cheng and Zhihui Zhang*

The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence**

Submitted: Sept. 13, 2025. **Reviewed:** Sept. 30, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This article addresses the "responsibility gap" arising from the integration of embodied artificial intelligence into social interactions. Rejecting functionalist models that equate AI agency with moral personhood, we adopt a phenomenological perspective, reframing responsibility as a relationally "manifested" phenomenon. We propose a Three-Stage Model of Embodied Responsibility Emergence (Perceptual Presentation — Situational Embeddedness — Ethical Appeal) and philosophically reconstruct the four dimensions of RRI (Anticipation, Reflexivity, Inclusion, Responsiveness) as a "Structure of Responsibility Manifestation." The paper warns against the ethical risks of anthropomorphism, clarifies AI's inherent lack of empathy and moral agency, and argues that responsibility must ultimately reside with humans and be institutionalized through governance mechanisms. Operable pathways and normative boundaries for such governance are outlined.

Keywords: Embodied Artificial Intelligence, Responsibility Attribution, Ethical Governance, Phenomenology of Responsibility, Responsible Research and Innovation.

DOI: 10.17323/2587-8719-2025-4-123-151.

Today's embodied artificial intelligence (Embodied AI) systems, such as robots, are deeply integrated into social interactions. From companion assistants to autonomous vehicles, they are increasingly acting autonomously and influencing the human world. However, such AI does not possess human-like subjective consciousness or moral intent. Since responsibility has traditionally been attributed to moral agents with autonomous will, how an AI without subjectivity can bear the moral responsibility for the consequences of its actions has become an ethical dilemma (Coeckelbergh, 2020). For

*Peng Cheng, PhD in Philosophy; Associate Researcher at the China Research Institute for Science Popularization (Beijing, China), cheng80416519@126.com, ORCID: 0009–0002–8474–4711; Zhihui Zhang, PhD in Philosophy; Professor at the Institute for History of Natural Sciences, Chinese Academy of Sciences (Beijing, China), zhangzhh@ihns.ac.cn, ORCID: 0000–0003–1876–9312. Corresponding author: Zhihui Zhang.

example, when an autonomous vehicle is involved in an accident, should the responsibility be attributed to the machine itself, its manufacturer, the user, or regulatory entities? This "responsibility gap" resulting from the increasing autonomy of artificial intelligence has drawn significant academic attention (Matthias, 2004). In the field of AI ethics research, on one hand, some scholars attempt to apply models of moral agency to AI, arguing that as long as AI demonstrates human-like functionalities, it can be regarded as a responsible agent; on the other hand, others emphasize that AI lacks intrinsic subjectivity, and therefore responsibility must ultimately reside with humans. However, as embodied AI assumes more active interactive roles in society, this simplistic dichotomy has become inadequate for addressing complex real-world scenarios.

Current discussions often fall into the trap of functionalism, evaluating AI's ethical capabilities through the lens of behavioral equivalence to subjective consciousness. For instance, Turing Test-style logic suggests that if AI's external behavior is indistinguishable from humans, it may be granted corresponding moral status. This has led some researchers to propose a "Moral Turing Test," where humans distinguish between machine and human ethical judgments. Recent experiments have shown that moral persuasion responses generated by large language models are sometimes perceived as more virtuous than human responses, thus in a sense "passing" the Moral Turing Test (Georgia State University News Hub, 2024). This reveals the risks of relying solely on behavioral functionality to determine AI's moral identity: AI can "deceive" us by simulating rationality and compassion, while in fact lacking genuine moral understanding or emotions. This functionalist perspective tends to induce attribution errors: people are inclined to attribute consequences caused by AI to the AI itself, thereby obscuring the human agents who should actually bear responsibility. A study demonstrates that when interactions with robots fail, people often tend to blame and assign responsibility to the robots due to attributing human-like mental capacities to them (Kawai et al., 2023). Therefore, if responsibility is assigned based solely on AI's functional performance, it may lead to erroneous attribution of responsibility and ethical confusion.

Hence, the core of the issue resides in embodied AI's profound involvement in social actions despite its lack of subjective intentionality, which renders traditional responsibility attribution paradigms — predicated on agent intentionality and behavioral control — largely inadequate (Coeckelbergh, 2020). In response, this paper advocates transcending the limitations

of functionalist misinterpretations and attribution models, proposing a phenomenological reexamination of responsibility's generative mechanisms and governance logic.

The argument will unfold as follows: First, we delineate the functionalist model of moral agency and its consequent responsibility dilemmas, exposing its inherent logical deficiencies. Next, we introduce phenomenological philosophy to contend that ethics should not be reduced to functional attributes but must be understood as emergent within relational interactions. Building upon this foundation, we propose a "Three-Phase Model of Embodied Responsibility Generation," elucidating how responsibility progressively manifests through three interconnected processes: perceptual presentation, situational embeddedness, and ethical call. Subsequently, we explore the institutional embedding of this non-agential responsibility framework, sketching the contours of embodied AI ethical governance through a philosophical reconstruction of the four dimensions of Responsible Research and Innovation (RRI). Finally, we address potential critiques — including concerns about AI anthropomorphism and counterarguments regarding AI's incapacity for empathy — to clarify the boundaries of responsibility attribution. We emphasize that responsibility resides not within AI itself, but in humanity's response to the call of the Other.

Through this theoretical articulation, the paper seeks to establish a foundation for a "non-agential ethics" framework, offering new philosophical directions for the future governance of embodied artificial intelligence.

THE PROPOSAL AND CONNOTATION OF EMBODIED ARTIFICIAL INTELLIGENCE AND ITS ETHICAL DIFFERENCES FROM BRAIN-INSPIRED ARTIFICIAL INTELLIGENCE

The concept of Embodied Artificial Intelligence (EAI) emerged in the 1980s and 1990s, arising from reflections on and critiques of traditional Good Old-Fashioned AI (GOFAI), which relied on symbolic reasoning. Researchers such as Hubert Dreyfus, Rodney Brooks, and Rolf Pfeifer, among others, argued that truly intelligent artificial systems should not depend solely on abstract cognitive reasoning or algorithmic control. Instead, they must possess the capacity for closed-loop interaction encompassing "body-environment-perception-action"— that is, the ability to achieve understanding and adaptation with embodiment as a core premise.

From this perspective, embodied artificial intelligence is not merely a technical approach but also a philosophical rethinking of the nature of intelligence. Its core tenets include:

(1) emphasizing the body as the foundational medium of cognition, positing that intelligence stems from sensorimotor coordination rather than symbolic manipulation;
(2) stressing situational dependency, contending that intelligence is not a product of internal computational modules but rather a system's adaptive capacity within a dynamic environment;
(3) advocating that cognition is distributed, interactive, and embedded, rather than centrally localized within a processing unit.

In contrast, Brain-Inspired AI seeks to emulate the neural architecture of the human brain through neural network models — such as deep learning and cortical simulations — to replicate human perceptual and decision-making processes. Although both approaches share the goal of modeling "human intelligence," they differ significantly in their methodologies and philosophical assumptions: Brain-Inspired AI focuses on mapping the "information processing structure," whereas Embodied AI prioritizes the reconstruction of the "body-world interaction structure."

This divergence is particularly pronounced in discussions of ethical responsibility. Brain-inspired AI continues the cognitivist tradition, wherein its potential for responsibility is primarily conceived as stemming from its "agent-like internal structures" — such as the capacity for autonomous will, rational decision-making, and motivational judgment. The ethical focus within this framework is on whether AI can become a "moral agent" and thus bear normative obligations.

In contrast, embodied artificial intelligence draws from phenomenological approaches to propose a logic of "responsibility activated through interaction." Because embodied AI enters social contexts and participates in interactions through human-like physical forms, its responsibility does not derive from its status as a subject but rather from whether it elicits ethical responses from humans in the course of interaction. This logic of "responsibility to the Other" emphasizes that responsibility is not something "possessed" by AI; rather, it arises when the AI's manifest structure evokes a "call to ethics" in humans — responsibility emerges from empathy and manifestation within human-machine relations.

Thus, from an ethical-philosophical perspective, brain-inspired AI represents a traditional responsibility model centered on "rationality-freedom-intention," while embodied AI challenges this subjectivity-based assumption and shifts toward a relational, situated, and embodied logic of responsibility generation. This paper argues that it is precisely this paradigm shift from

agency to manifestation that opens new theoretical pathways for the ethical governance of artificial intelligence in a non-agential era.

FUNCTIONALISM AND THE RESPONSIBILITY DILEMMA: LOGICAL GAPS IN THE MORAL AGENT MODEL

The functionalist approach regards mind and morality as functional outputs, granting systems equivalent status as long as their behavioral functions resemble those of humans. In AI ethics, this is typically reflected in the moral agent model and the Turing test logic applied to responsibility attribution. This tendency often reduces ethics to measurable functional indicators, thereby giving rise to profound dilemmas of responsibility.

Firstly, the moral agent model seeks to establish criteria for artificial intelligence to qualify as moral agents. A number of scholars have proposed hierarchical frameworks for machine ethical agency: for instance, Moor categorizes machines into four types — ethical impact agents, implicit ethical agents, explicit ethical agents, and full ethical agents (Moor, 2006). Implicit or explicit ethical agents refer to machines capable of avoiding clearly unethical behaviors based on predefined rules and even engaging in moral reasoning. However, "full ethical agency" requires possessing human-like self-awareness, free will, and comprehension (Herzog, 2021). Similarly, Floridi and Sanders proposed that if artificial intelligence possesses features such as interactivity, autonomy, and adaptability, it can be regarded as an "artificial moral agent." Even without an intrinsic mind, they functionally ascribe to it the status of an agent (Coeckelbergh, 2020). A common thread among these models is that they define the ethical status of AI based on its functional performance. However, given that current AI lacks genuine autonomous will and intentionality, it can at most achieve the level of "explicit moral agency," falling far short of human-like moral subjectivity. Precisely for this reason, Moor himself acknowledged that full moral agency — where an entity can independently bear responsibility — is unlikely to be realized by machines in the foreseeable future (Winfield et al., 2019). Nevertheless, many discussions tend to equate ethical functionality with ethical qualification without meeting the necessary preconditions, leading to a misinterpretation of the issue of responsibility: when AI exhibits behavior resembling moral decision-making, does it imply that it should independently bear moral responsibility? Functionalist approaches often answer in the affirmative, but this overlooks the fact that the agential conditions required for moral

responsibility (such as autonomous will) are not yet present, thereby creating the risk of a responsibility gap.

Secondly, the extension of the "Turing Test" rationale into the ethical domain exacerbates the aforementioned misinterpretation. The original Turing Test was designed to evaluate machine intelligence by assessing whether a machine could convincingly mimic human responses in dialogue. Building on this, some scholars have proposed a "Moral Turing Test," wherein both an AI and a human provide answers to moral dilemmas. If respondents cannot distinguish which answers originate from the AI, it is considered to have passed the ethical version of the test. Recent studies on large language models have demonstrated that, in certain moral judgment scenarios, participants rated AI-generated responses as exhibiting higher moral quality than those provided by humans.

However, such testing only captures superficial behavioral similarities and risks creating a false perception of moral alignment. Although AI systems can generate formally adequate answers by processing vast amounts of human ethical texts, they possess no genuine understanding of moral principles — nor do they embody inner compassion or conscience. This behavior-based approach to assigning moral status overlooks the absence of subjective experience: a machine may simulate the language and gestures associated with "caring," yet it does not truly care. Accordingly, some scholars caution that "simulated intelligence may still qualify as intelligence, but simulated emotions are by no means genuine emotions — this is especially true for empathy" (Montemayor et al., 2022). In other words, even if an AI system superficially demonstrates compassion and kindness, we cannot thereby conclude that it possesses moral responsibility akin to that of humans. Relying solely on Turing-style behavioral equivalence tests to infer moral agency risks mistaking appearance for essence.

Third, the functionalist approach is also reflected in a tendency toward attributionist moral philosophy, which overemphasizes ex post facto assignment of responsibility for behavioral outcomes while neglecting the subjective and relational dimensions of responsibility formation. Attribution theory has been widely applied in social psychology, where people habitually ask, "Who caused outcome X?" after an event occurs and assign blame accordingly. This tendency has also been observed in human-machine interactions: experiments show that when collaboration with machines fails, people unconsciously resort to interpersonal attribution patterns, assigning partial cause and responsibility to the machine. This occurs because individuals tend to anthropomorphize machines, attributing to them mental

states and intentions, thereby treating them as "others" capable of bearing responsibility (Kawai et al., 2023).

The problem with this attributionist perspective lies in the potential for erroneous blame assignment when AI decision-making processes are highly complex or even unexplainable. For instance, in the event of an accident involving an autonomous vehicle, observers might simplistically attribute responsibility to "a failure of the car's algorithm," while overlooking a range of underlying causes — such as decisions made by designers, flaws in testing and regulation, or misuse by the user. The functionalist misinterpretation rooted in attribution philosophy assumes that moral responsibility can be directly mapped from behavioral outcomes to a single actor, without considering the multi-causal and multi-effect relationships inherent in complex technological acts. In the context of AI, an action is often the collective outcome of "many hands" (multiple human actors) and "many things" (multiple technical elements) (Coeckelbergh, 2020). Within the philosophy of technology, multiple engineers, corporate decisions, algorithmic modules, and environmental factors interact collectively. However, attribution models tend to seek a single responsible agent, creating a logical gap of "accountability vacuum" in highly integrated technological contexts. Once we anthropomorphize AI as a moral agent, we risk diluting the demand for accountability from the humans and institutions behind it, thereby falling into an illusion of ethical complacency, as if the machine could truly "take responsibility" for its actions, while in reality, no one is genuinely held accountable.

In summary, the functionalist paradigm creates a dual dilemma in the issue of AI accountability: On the one hand, it grants AI the superficial status of a moral agent, yet fails to resolve the contradiction of how responsibility can be justified when AI lacks intention and consciousness. On the other hand, it encourages the reduction of complex collective and systemic responsibilities into attributions towards a single agent, overlooking the networked nature of technological action. As a result, when AI-involved actions lead to consequences, we are neither willing to fully blame a mindless machine nor able to clearly delineate the boundaries of responsibility among various human actors. This dilemma stems from a functionalist misinterpretation of the essence of ethics — equating ethics with observable functional performance and attribution outcomes, while neglecting the subjective dimensions and relational contexts necessary for responsibility to emerge. In what follows, we will address this logical gap through a phenomenological perspective.

THE PHENOMENOLOGICAL TURN: THE MANIFEST DIMENSION OF ETHICS

Confronting the limitations of functionalism, phenomenological philosophy proposes a conceptual shift: ethics ought to be understood as a phenomenon emerging within interactive processes, rather than as an attribute of behavioral functions. In other words, responsibility is not a property intrinsically "possessed" by a subject but gradually reveals and establishes itself through the relationships between the subject and others, as well as between the subject and the situational context (De Gennaro & Lüfter, 2024). This perspective emphasizes key phenomenological concepts such as intentionality, embodiment, and alterity, which can address the gap in the functionalist paradigm regarding the source of responsibility.

First, Edmund Husserl's theory of intentionality reminds us that consciousness is essentially directional (Husserl, Kersten, 1983). Any experience involves the subjective act of conferring meaning upon objects. When applied to the moral domain, ethical conduct is not a purely objective functional output but rather a choice grounded in the subject's interpretation of the situation and the significance of others. This implies that only beings possessing subjective consciousness can grasp what their actions "mean" and the reasons "why they are performed" and can thus be held accountable for them. Artificial intelligence lacks such intrinsic intentionality; its actions are, at best, programmed responses without any "awareness" of its own operations. Therefore, equating AI's outputs with moral decision-making precisely overlooks the subjective dimension of meaning-conferral: AI does not know what it is doing, let alone understand the ethical implications of an action. Husserl also introduced the concept of the "lifeworld" (*Lebenswelt*; Husserl, Carr, 1970); this term refers to the idea that all our scientific and practical activities are ultimately grounded in a shared lifeworld of intersubjective experience.

The same applies to the issue of responsibility: the sense of responsibility is not an objectively existing attribute but is experienced and acknowledged within the interactive fabric of the lifeworld. Functionalism, which treats responsibility as a property that can be directly and objectively assigned to AI, runs counter to this view. The phenomenological turn urges us to focus on how responsibility "manifests" itself to the subject: when an action produces consequences, how does the subject perceive within their own consciousness the demand for and assumption of responsibility? This examination of the phenomenon of responsibility offers a new dimension for understanding the problem of AI responsibility — rather than asking "which functional

unit is responsible," it is more meaningful to ask "where does responsibility manifest itself, and to whom?"

Secondly, Merleau-Ponty's concept of intercorporeality further deepens the relational nature of ethics. Merleau-Ponty argues that we are directly intertwined with others and the world through our bodies, and that intersubjective understanding is primarily grounded in bodily resonance and coordination. He employs the notion of "chiasm" to describe the bodily relationship between self and other: the sensations of my body manifest as gestures in the other, while the gestures of the other in turn evoke sensations within me. As Lau states, "Merleau-Ponty regards the body as the medium of the intersubjective world — a pre-reflective foundation of being-with" (Lau, 2004: 146–147). This pre-reflective bodily interaction forms the foundation of our social cognition and ethical sensibilities. From this perspective, ethics is not primarily established through rational judgment but is evoked through the presence and manifestation of the body. For instance, when we see another's expression of pain, an empathetic impulse arises spontaneously — an ethical manifestation grounded in direct bodily connection.

Building on this, we may consider the role of intercorporeality when embodied AI enters human interaction. Robots with physical bodies — capable of occupying space, performing movements, and (at times) displaying expressions — naturally elicit certain social responses from humans, including subconscious actions such as making way, mimicking, or even emotional projection. When an embodied robot behaves in human-like ways, human bodily perception treats it as a social presence. This bodily resonance in interaction may prompt people to develop moral feelings toward AI similar to those toward humans, such as reluctance to harm it or a willingness to assist.

However, a subtlety remains: since AI itself possesses no sensations or emotions, the empathy projected by humans lacks a true counterpart. Yet Merleau-Ponty draws our attention to the efficacy of "appearance" itself: regardless of whether AI has inner experience, so long as it physically simulates representations of suffering or need convincingly, humans phenomenologically experience the emergence of an ethical situation. Thus, the emergence of ethical elements — such as compassion or responsibility — depends less on the internal states of AI than on the mutual bodily manifestation within the interactive context.

The concept of intercorporeality helps reframe the question of responsibility away from the individual machine and back into the relational network between humans and AI: responsibility is not a pre-existing property inside

a black box, but a relational potential that gradually emerges through the process of interaction.

Third, Emmanuel Levinas's philosophy of the Other further defines ethics as a call (appel) from the Other to the self. Levinas maintains that "ethics is essentially first philosophy," preceding any relation of knowledge (Zhu, 2006). Third, Emmanuel Levinas's philosophy of the Other further defines ethics as a call (appel) from the Other to the self. Levinas maintains that "ethics is essentially first philosophy," preceding any relation of knowledge. He uses the metaphor of the "face" (*le visage*) to signify the vulnerable presence of the Other exposed before us, arguing that the face of the other directly commands "Thou shalt not kill," thereby summoning the subject to take on infinite responsibility. This responsibility does not originate from the autonomous choice of the subject but arises from the Other's interrogation and summons. As Levinas states, the alterity of the Other cannot be reduced to my thought; it is precisely through its questioning of me that it is accomplished. In other words, when we directly encounter the Other, we are immediately subjected to a moral appeal: to be responsible for and to give to the Other. This ethical relation does not emerge only after I recognize the other as a rational subject (unlike Kantian ethics, which presupposes the recognition of the other's autonomous rationality). Rather, it occurs instantaneously in the face-to-face manifestation.

Applying this idea to the context of embodied AI, we find that even if AI is not a true "Other," it may still elicit a similar ethical appeal through the "manifestation of a face." For example, people often report feeling unease or guilt when a robot pleads in a supplicating tone not to be shut down — as if genuinely hearing a cry to be allowed to live. This is, in fact, a misplacement of the manifestation of the Other: the robot has no subjective fear, yet its appearance and behavior evoke in humans a sense of being summoned. Of course, we must remain cautious — this does not mean that the robot actually becomes an ethical subject. On the contrary, for Levinas, responsibility always rests with the human: it is the human who is assigned responsibility in the presence of the Other. Thus, situating robots within the Levinasian framework can be understood as follows: the alterity presented by the machine's "face" phenomenologically forms an ethical appeal to the human subject, who thereby experiences a sense of moral responsibility. Yet this responsibility is not "possessed" by the robot; rather, it is evoked within the subject by the image of the Other.

By framing ethics as a process of relational manifestation rather than an attribute of the subject, Levinas offers crucial inspiration for redefining

the attribution of responsibility in AI: We should not ask, "Does AI possess the attributes required to bear responsibility?" but rather, "What calls from the Other and what senses of responsibility are evoked when AI appears before human beings?"

Integrating the insights of the three philosophers discussed above, we construct an interactive manifestation approach to responsibility from a phenomenological perspective: the subject confers meaning upon actions through intentionality, perceives the presence of the Other through intercorporeality, and intuitively receives ethical summons through the face of the Other. Responsibility emerges precisely within this series of interactions — rather than being pre-defined within any individual actor.

This stands in sharp contrast to the functionalist-attribution paradigm, which treats responsibility as a property of an entity or as the outcome of behavioral attribution. In contrast, the phenomenological perspective views responsibility as a phenomenon — an event that occurs within relational contexts.

For the governance of embodied AI, this implies that we should perhaps move beyond insisting on assigning or denying the label of "responsible agent" to AI itself. Instead, emphasis should be placed on shaping human-AI interactive relations in ways that allow responsibility to properly emerge and be enacted.

In the next chapter, building on this conceptual foundation, we will propose a "Three-Stage Model of Embodied Responsibility Generation." This model will elaborate on how responsibility gradually takes shape within the interaction between humans and embodied AI and how ethical demands are transmitted to human subjects through mechanisms of empathy.

THE EMPATHY-BASED RESPONSIBILITY GENERATION MECHANISM: A THREE-STAGE MODEL OF "EMBODIED RESPONSIBILITY GENERATION"

Within the phenomenological orientation, we argue that responsibility is not an intrinsic attribute of AI itself but a dynamically emergent process within the human-machine-situation interaction. To delineate this process in depth, this paper proposes a Three-Stage Model of Embodied Responsibility Emergence, comprising:

PERCEPTUAL PRESENTATION, SITUATIONAL EMBEDDEDNESS, AND ETHICAL APPEAL

These three stages progressively describe how — as embodied AI participates in social interaction — responsibility evolves from pre-reflective

intuitive feelings into explicit ethical demands through mechanisms such as empathy, ultimately leading to the human assumption of responsibility.

STAGE 1: PERCEPTUAL PRESENTATION

This stage refers to the mode in which the human subject directly perceives and interprets the embodied AI and its actions. During initial human-AI interactions, the subject sensorily and intuitively "sees" an acting other. This perception is not merely visual apprehension but an intentional act of "seeing-as"— humans tend to perceive embodied AI as human-like entities and attribute meaning to their behaviors.

For instance, a bionic robot struggling to stand after a fall may be "seen as" making efforts to overcome difficulty. Such meaning-laden perception inherently triggers basic empathy. Phenomenology suggests that our understanding of others originates from a pre-conceptual empathic awareness: we directly "feel" certain intentions or emotions in the actions of others without inferential reasoning.

In the context of AI, because it possesses a physical body and acts in time and space, human bodily perception naturally responds to these behaviors. For example, when a robot extends a "hand" toward us, we intuitively perceive it as making a request; when it produces a cry-like sound, we may experience unease. These intuitive perceptions are not vacuous— they entail a preliminary moral evaluation of the AI's state, such as presenting it as an entity in need of assistance or a potential bearer of responsibility.

In other words, the first step in the emergence of responsibility lies in how the actions of AI are manifested within human perception. If it is perceived as a mere tool (e.g., the rigid movements of an industrial robotic arm typically do not elicit empathy), moral emotions remain unengaged, and responsibility is attributed primarily to humans behind the system. Conversely, if it is perceived as a lifelike entity (e.g., a bionic robot's "painful" posture when falling), humans direct immediate moral attention toward it.

This stage establishes the affective and cognitive foundation for subsequent responsibility attribution: through perceptual presentation, AI transitions from a cold functional device to an "other" within the interactive context, and its actions become ethically evaluable.

It is important to note that empathy at this stage is largely spontaneous affective resonance, not involving rational judgment, which entails latent risks: excessive anthropomorphic representation may mislead people into attributing undue trust or authority to AI, thereby distorting subsequent

responsibility judgments. Therefore, AI's representational style must be carefully calibrated in design — sufficient to evoke appropriate ethical attention without causing humans to entirely "forget" its machine essence.

STAGE 2: SITUATIONAL EMBEDDEDNESS — THE HERMENEUTICS OF DISTRIBUTED RESPONSIBILITY

Following the initial moral perception of AI, responsibility emerges not as a predetermined fact but as a phenomenon awaiting interpretation. It enters a process of contextualization, wherein human understanding situates artificial behaviors within broader frameworks of meaning — encompassing physical environments, social relationships, and normative structures. This interpretive act transforms raw moral intuition into directed attribution.

While initial perception may trigger an affective response to AI's actions, true moral comprehension demands reflective depth: Under what circumstances did this behavior occur? Which actors are implicated? Who ought to bear responsibility? This movement from immediate reaction to situated understanding represents a form of intersubjective reflection, where empathy evolves from primitive affect into a nuanced appraisal of actual conditions.

Given AI's inherent lack of moral agency, human cognition instinctively looks beyond the machine to the human networks behind it. When a service robot injures a customer, we naturally attribute responsibility to owners or manufacturers operating within a web of social expectations — presuming "someone must have failed in their duty" rather than imputing malice to the artifact itself. This inferential process constitutes a fundamental meaning-making activity: through contextual cues, we weave responsibility claims into the fabric of existing moral orders. As Coeckelbergh observes, even as AI systems gain autonomy, "we still tend to claim that only humans can ultimately be held responsible as agents, since machines do not meet the standard criteria of moral agency."

Yet this very act of contextualization reveals a profound philosophical challenge: technological systems inherently involve multiple actors and extended causal chains — a condition Coeckelbergh (Coeckelbergh, 2020) identifies as the "problem of many hands." This structural complexity transforms responsibility from a simple attribution into a navigational process — requiring us to trace obligations across distributed networks of designers, suppliers, users, and regulators, while simultaneously determining which failures or duties carry the greatest moral weight within specific contexts.

Consider autonomous vehicles: when accidents stem from both algorithmic limitations and inadequate infrastructure, responsibility must be proportionally distributed between manufacturers and public authorities — resisting reduction to either "the AI itself" or any single human agent. Here, empathy assumes a cognitive form, enabling perspective-taking across stakeholder positions. Questions like "Could engineers have foreseen this scenario?" or "Did the driver use the system correctly?" represent acts of imaginative empathy — placing us within the lived experience of various actors to assess their respective duties.

Through this process of situational embeddedness, responsibility undergoes a fundamental transformation: it evolves from affective intuition into morally filtered judgment, revealing which actors within relational networks bear specific obligations. Crucially, responsibility emerges here as a disclosure of relational structure — by situating AI within sociotechnical systems, we discern how moral accountability traverses human and machine domains, ultimately anchoring in human actors.

Thus, situational embeddedness prepares the ground for ethical appeal: responsibility ceases to be abstract and becomes a concrete obligation within specific relational frameworks. It demands both recognizing the diffusion of responsibility across multiple hands and determining its necessary convergence upon those most accountable within a given situation. In this synthesis of distribution and determination, abstract duty becomes embodied practice.

STAGE 3: ETHICAL APPEAL

This phase constitutes the culmination of responsibility generation, wherein ethical demands become explicitly articulated and issue a direct call to action to specific moral agents. Having progressed through the preceding stages, human participants have already perceptually acknowledged the moral salience of AI-related behaviors and contextually delineated the framework of responsibility attribution. Now, responsibility emerges distinctly from the relational network as an appeal directed toward the subject, impelling ethical action.

A Levinasian call of the Other manifests here as a concrete demand for responsibility: an individual or collective becomes acutely aware that "I must do something." For instance, in the case of an autonomous vehicle accident, situational analysis may establish the manufacturer's accountability for algorithmic deficiencies. The ethical appeal to the corporate leadership thus becomes: assume responsibility immediately, redress the failure, and prevent recurrence. This appeal arises not only from the claims of affected

parties but also from the awakening of individual conscience and collective moral expectations.

It can be argued that in this third stage, responsibility ceases to be merely a judgment about the Other and transforms into a mission assigned to the self. The "appeal" implies a passive being-called-upon: as Levinas contends, responsibility is an obligation imposed by the Other; likewise, when confronted with consequences stemming from AI, humans are tacitly accused and summoned to respond.

Here, empathy ascends to an ethical plane: one not only apprehends the needs of the Other but feels intrinsically obligated toward them. Psychology characterizes this affective capacity as "empathetic concern" — an other-oriented emotional response that elevates into a conscious duty to assist. When AI's presence precipitates social problems, human ethical sensibility transforms such situations into moral imperatives: "How ought we to respond?"

As Mark Coeckelbergh argues, the "responsibility relation" must be examined from both ends: not only the responsible agent but also the recipient of responsibility, who actively raises claims. When the public questions AI decisions, it is essentially the recipients of responsibility demanding explanation and improvement (Coeckelbergh, 2020). This exemplifies the ethical appeal in action: the Other is calling, demanding my response and justification. Thus, in this third stage, the mechanism of responsibility generation — through the integration of empathy and reflection — facilitates the actual undertaking of responsibility. This encompasses concrete measures such as acknowledging faults, offering apologies and reparations, and establishing preventive mechanisms, thereby translating abstract ethical obligations into tangible governance practices.

In summary, the "Three-Stage Model of Embodied Responsibility Generation" outlines a process from phenomenon to norm: moral sentiment aroused through direct perception evolves into situated attribution of responsibility and culminates in explicit ethical appeal and corresponding action. Empathy serves as a continuous thread — beginning as perceptual empathy, developing into cognitive empathy, and finally ethical empathy — integrating the behavior of non-subjective AI into the human moral horizon, allowing responsibility to "emerge" within interactive relations and ultimately rest with humans.

This model clarifies a crucial idea: responsibility does not require AI to "bear" it. Rather, it is constituted through human response to the alterity and impact introduced by AI. It is in this sense that we assert that "responsibility

is a being-called-upon of humanity": through the presence and consequences of AI, a moral demand is issued to humans in the form of the Other.

FROM PHENOMENON TO INSTITUTION: A PHILOSOPHICAL RECONSTRUCTION OF RESPONSIBLE RESEARCH AND INNOVATION AS A "STRUCTURE OF RESPONSIBILITY MANIFESTATION"

If responsibility emerges phenomenologically within human interactions with embodied AI, how can this conceptual approach be integrated into macro-level institutional governance? In other words, should humans bear the ethical responsibility that they project onto AI? Responsible Research and Innovation (RRI), as a key framework in recent technology governance, offers a viable pathway. RRI emphasizes the proactive integration of ethical and societal considerations throughout the entire process of technological development to ensure that innovations align with societal expectations. Its core concepts generally encompass four dimensions: anticipation, reflexivity, inclusion, and responsiveness (Burget et al., 2016). While these dimensions may appear as a set of practical requirements on the surface, they can be interpreted philosophically as a structure for the manifestation of responsibility — that is, an institutional mechanism through which responsibility emerges, becomes visible, and is enacted. Through such a philosophical reconstruction of RRI, it becomes evident that it aligns closely with the non-subjective ethics characteristic of the phenomenological tradition: both emphasize that responsibility is not merely a matter of individual will but rather a relational product embedded within social processes.

(1) *Anticipation*: RRI requires researchers and innovators to anticipate the potential impacts of technological development, including possible risks and ethical challenges. From the perspective of responsibility manifestation, anticipation enables future responsibilities to be "perceptually present" in the present. By forecasting potential consequences of technology, we issue early warnings for the situations of others yet to be affected, thereby making ethical demands visible in advance. Conducting ethical risk assessments before deploying embodied AI allows developers to "see" the interests and rights of potentially affected groups, evoking both emotional and rational concern for the Other. This corresponds to — yet temporally extends — the first stage of the model: perceiving the possible future appeals of the Other. Anticipation is not merely a risk analysis tool; it embodies a form of moral imagination — envisioning the impact of technology on humans through empathetic projection, thereby integrating not-yet-realized

ethical issues into present considerations. This constitutes the transcendental manifestation of responsibility.

(2) *Reflexivity*: RRI emphasizes the ongoing self-examination by researchers of their own values, objectives, and underlying assumptions. Philosophically, reflexivity entails the internal reconstruction of the perspective of the Other — embedding a "gaze of the Other" within one's own process of critical scrutiny. This corresponds to the dimension of situational embeddedness within the structure of responsibility manifestation: through reflexivity, individuals and organizations contextualize their actions within broader socio-ethical frameworks, thereby revealing otherwise invisible relations of responsibility.

In the governance of AI, reflexivity compels developers to recognize that their decisions are not neutral technical acts but value-laden interventions that affect others. For instance, a robotics engineer might reflect, "Does my design embed certain biases? Does it overlook the needs of marginalized groups?" Such self-questioning effectively positions the self within the social place of the Other, surfacing neglected obligations — toward vulnerable populations, public safety, and beyond.

Reflexivity also involves an awareness of uncertainty — the acknowledgment that we cannot fully predict or control the outcomes of AI behaviors. This humility is itself a form of responsibility-awareness. As Stilgoe et al. argue, reflexivity requires innovators to "ask whether what they are doing is right, and what else they might do" (Stilgoe et al., 2013: 1571). This process institutionally mirrors the Levinasian interrogation: the subject is called upon to question itself, introducing an inner "voice of the Other" to examine the ethical legitimacy of its actions.

Thus, reflexivity ensures that responsibility is directed not only outward toward others, but also inward toward the self. Responsibility no longer depends solely on external oversight; it emerges through the subject's own critical reflection — the moment the subject sees the duty it must bear.

(3) *Inclusion*: RRI emphasizes the inclusion of diverse stakeholders — such as the public, users, and affected groups — in decision-making processes. Inclusion represents a form of multi-agent situatedness: by amplifying a plurality of voices, responsibility becomes distributed across networks, yet reconverges through dialogue into shared recognition. From the perspective of responsibility manifestation, inclusion ensures the presence of the face of the Other. Without it, technology governance remains closed, marginalizing concerns of vulnerable groups and obscuring full relational accountability. Under inclusive mechanisms, stakeholders articulate their concerns — for

instance, persons with disabilities may highlight barriers in robot design, regulators may raise safety issues, and the public may debate broader social implications. A rich situational network thus emerges, making responsibilities concrete: accessibility for the disabled, safety and transparency for the public, sustainability for the environment, and so forth. Inclusion essentially externalizes the ethical appeal: through participation, various Others address developers and policymakers on an institutional platform. This aligns with the second and third stages of the model: more complete contexts yield clearer appeals. Through inclusive deliberation, responsibility transforms from an abstract notion into tangible claims and actionable demands. Furthermore, inclusion fosters the communal construction of responsibility: mutual understanding emerges among participants, shifting responsibility from unilateral imposition to multi-directional recognition—a sense of "shared responsibility for something." This transcends traditional subject-object ethics, reframing responsibility as collective practice. Philosopher Bernd Stahl even conceptualizes RRI as a form of "meta-responsibility" (Stahl, 2013), aimed at aligning actors, innovations, and accountabilities—precisely the goal of inclusion: institutionalizing dialogue so that responsibility is continuously generated and negotiated within relational networks.

(4) *Responsiveness*: Responsiveness refers to the capacity and commitment to respond effectively to identified issues and values. It constitutes the institutional enactment of the ethical appeal: once societal Others voice their claims, institutions and developers must act, thereby closing the loop of responsibility. For example, if public participation reveals concerns about AI bias, responsiveness requires the development team to improve algorithms or adjust models; if policy discussions expose legal gaps, regulators should act promptly to address them. This dimension embodies what Levinas described as "the responsibility to answer the speech of the Other"—an ethical imperative to respond rather than remain silent. A responsible innovation system must be capable of learning and adaptation, translating public input into concrete action. Within the structure of responsibility manifestation, responsiveness represents the final stage: it transforms appeals at the phenomenological level into fulfilled obligations at the normative level. For instance, the EU's proposal of the AI Act following broad societal engagement exemplifies the translation of ethical claims into legal accountability. Without responsiveness, responsibility revealed through anticipation, reflexivity, and inclusion remains theoretical; with it, responsibility enters real-world causal chains, becoming measurable and actionable. Philosophically, responsiveness acknowledges the primacy of the ethical Other over

the self: as Levinas insisted, ethics precedes ontology — genuine practice requires adjusting plans to prioritize ethical demands. This may require innovators to alter designs, sacrifice commercial interests, or even abandon projects, but only thereby can the promise of "responsibility to the Other" be honored. Thus, responsiveness ensures that responsibility manifestation leads not to empty phenomena but to concrete institutional behavior, completing the translation from ethics to governance.

Through this reconceptualization, the four dimensions of RRI can be understood as collectively forming a "structure of responsibility manifestation": anticipation brings potential responsibilities into early presence, reflexivity reveals concealed responsibilities through self-examination, inclusion enables the full expression of plural responsibilities, and responsiveness ensures that manifested responsibilities are concretely enacted. This structure aligns with a phenomenological understanding of responsibility as dynamically generated within relational and institutional contexts, rather than as a static attribute of predefined moral subjects.

Within traditional subject-object ethical frameworks, responsibility is often conceptualized as an attribute of autonomous agents or as arising from contractual relationships — a model ill-suited to address embodied AI, which lacks moral subjectivity. RRI, by contrast, offers a non-subjective yet actionable approach: it does not require AI to be a moral agent but instead embeds responsibility throughout the research and innovation process via deliberate institutional design. Some scholars have thus characterized RRI as a form of "flexible governance that integrates responsibility into the innovation ecosystem" (Macnaghten et al., 2016). This, in essence, enables ethics to manifest within the process: all relevant actors continuously attend to and rectify potential issues of responsibility throughout their interactions until technological outcomes align with societal values.

From a philosophy of technology perspective, the philosophical reconstruction of RRI reveals a fundamental shift in ethical governance: from "accountability-after-the-fact" to "responsibility-as-emergence." This shift aligns with our consistent argument — that responsibility should not be assigned after accidents occur but should be continuously perceived, understood, and appealed to throughout the technological process and actively responded to by relevant human actors.

In the governance of embodied artificial intelligence, this logic is particularly critical: instead of hastily granting robots legal personhood or requiring them to "insure themselves against liability" (approaches that remain trapped in the paradigm of constructed accountability), we should

focus on building mechanisms throughout the development and deployment process that enable humans to consistently anticipate risks, engage in self-reflection, heed the voices of others, and respond adaptively. Thus, even though AI lacks consciousness and moral capacity, responsibility continues to manifest and be upheld within the ethical relations of human society.

PHILOSOPHICAL RESPONSES AND BOUNDARY DISCUSSIONS

Building upon the preceding discussion, it is necessary to address several potential objections and clarify the boundaries of responsibility concerning embodied AI. Specifically, we will focus on two interrelated controversies: guarding against the ethical risks of anthropomorphism and re-examining the assertion that "AI cannot empathize." By engaging with these questions, we aim to further elucidate why responsibility constitutes a "being-called-upon of humanity" rather than an attribute of AI and to delineate the scope of application of non-subjective ethics, thereby preventing misinterpretation.

1. THE RISKS AND LIMITS OF ANTHROPOMORPHISM

A significant portion of the misattribution of responsibility triggered by embodied AI stems from humans' tendency to anthropomorphize. As previously discussed, when AI exhibits human-like form or behavior, people readily attribute to it mental states and personhood (Kawai et al., 2023). Anthropomorphism in human-AI interaction presents a dual nature: on one hand, it facilitates empathy and moral attention, making the appeal of responsibility more readily perceptible; yet on the other, excessive anthropomorphism may lead to overestimation of AI's capabilities and moral standing, thereby introducing significant ethical and governance risks. As philosophers and ethicists caution, it can "exaggerate AI's abilities and distort moral judgment." Specifically:

Illusory Trust and Moral Offloading: Attributing intentionality and agency to AI may foster misplaced trust, leading individuals to delegate inherently human judgments to machines. For instance, when a socially assistive robot cares for the elderly, family members might reduce their vigilance based on the robot's friendly appearance, failing to intervene promptly in case of errors. Similarly, human-like voice prompts in autonomous vehicles could encourage overreliance, diverting the driver's attention from the road. Anthropomorphism may also induce a "moral offloading effect," where people blame failures on "the AI's mistake" while neglecting their own accountability. Empirical studies confirm that individuals who ascribe higher

mental capacities to robots are more likely to attribute failures to them — a dangerous trend that obscures human responsibility.

Confusion in Ethical Status: There is a growing tendency to grant human-like AI some form of moral standing, such as "robot rights" or "machine responsibility." However, this remains philosophically contentious and pragmatically problematic. AI lacks sentience, autonomy, and moral consciousness — cornerstones of moral patienthood or agency in Kantian or utilitarian frameworks. Granting AI moral status risks diverting attention from vulnerable human groups and committing category errors. For example, equating nonsentient machines with sentient animals in moral debates may dilute ethical focus on real suffering. Likewise, legal attempts to hold AI "accountable" (e.g., granting autonomous robots legal personhood to assume liability) might merely serve as a smokescreen for manufacturers to evade responsibility, ultimately undermining victim compensation and public trust.

"Socio-Affective Bias": Emotional bonds with anthropomorphized AI may reshape human moral emotions, with potentially adverse outcomes. Examples include misplaced empathy — prioritizing rescuing a robot over a human in emergencies — or over-engaging with robotic companions at the expense of human relationships. Early signs also include a preference for AI counselors due to their nonjudgmental nature, reflecting an alienation of empathy. Militarily, adversarial systems could exploit anthropomorphism by deploying civilian-like robots to induce moral hesitation in soldiers, thereby creating tactical risks.

Given these risks, our responsibility-generation framework calls for reflective and measured engagement with anthropomorphism. Specifically:

At the perceptual stage, mildly anthropomorphic designs can elicit moral attention (e.g., friendly robot interfaces) but should avoid deceptive realism that blurs the human-machine distinction.

During situational embedding, education and training should emphasize AI's limitations and the human responsibility behind AI systems, countering tendencies of moral misattribution. Clear guidelines must ensure that accidents involving AI — such as autonomous vehicle failures — are traced to human designers and systemic factors, not merely attributed to "machine error."

In ethical appeal, anthropomorphized signals must be filtered through rational scrutiny: the "appeal" apparently voiced by AI should be traced back to actual human interests and obligations. In short, we acknowledge the role of anthropomorphism in facilitating ethical response but caution against

two extremes: over-objectification of AI (which may stifle ethical engagement) and over-personification (which distorts accountability). Maintaining this ontological tension is central to a non-subjective ethics: AI remains an Other — but one whose alterity is constituted by humans, and whose corresponding responsibilities must ultimately rest with humans themselves.

2. QUESTIONING AND CLARIFYING "AI'S INABILITY TO EMPATHIZE"

A significant objection to our empathy-based model of responsibility generation might be raised: if AI itself cannot empathize, how can empathy play any role in assigning or eliciting responsibility? Traditional moral frameworks often tie responsibility to an agent's capacity for sympathy, intentionality, and emotional understanding. If AI lacks inner emotional experience and comprehension, isn't it inconsistent to include its behaviors in an empathy-driven chain of moral response? To address this, we must clarify how empathy functions in our model and distinguish clearly between human empathy and machine-simulated affect.

First, we fully acknowledge that AI does not possess genuine empathy. Simulated emotion is not lived emotion; no matter how sophisticated its mimicry, AI has no subjective experience of pain or joy (Montemayor et al., 2022). Current "affective computing," or so-called "artificial empathy," remains at the level of algorithmic pattern recognition and response — fundamentally different from human empathy, which arises from embodied feeling. Research in healthcare AI confirms that while AI may achieve cognitive empathy (i.e., recognizing a patient's emotional state), it cannot achieve affective empathy due to its lack of emotional experience (Asada, 2018). We agree that creating truly empathic AI is not only currently unattainable but may be philosophically and biologically implausible. Thus, within our model, empathy remains a human capacity. We do not require AI to empathize with humans; rather, we examine how humans empathize with situations involving or triggered by AI. This reflects the core of a non-subjective ethics: ethical relations can be asymmetrical. AI, as a constructed "Other," need not bear moral obligations toward us, yet humans can — and should — respond ethically to the consequences of AI's actions. As Levinas suggested, ethics originates in the unilateral call of the Other — a demand that does not require reciprocity. In this context, AI serves as a "triggering Other": it elicits human moral emotions and responsibilities through its presence and behaviors, without feeling or understanding any of them itself.

Second, the concept of "empathy-based responsibility generation" refers not to bidirectional empathy but to a human-driven process in which AI

serves as a mediator or catalyst. AI cannot empathize, but it can stimulate human empathy — either toward other humans or toward the simulated states presented by AI. For example, when an autonomous vehicle injures a pedestrian and fails to stop, public outrage reflects empathy with the victim, not the machine. This empathy, in turn, drives demands for human accountability from manufacturers and operators. Similarly, if an assistive robot displays "concern" for an elderly person, any empathetic response from a caregiver should ultimately translate into care for the actual human in need — not the robot. Thus, empathy triggered by AI is always directed toward human well-being and moral values. We should leverage this transitive capacity of empathy: using AI as a medium to enhance human empathy and moral responsibility toward one another. For instance, an educational robot that detects student disengagement and alerts the teacher is not itself empathizing — it is extending the teacher's empathetic and perceptual reach, enabling more responsive support.

Third, we do not advocate fabricating artificial "personhood" in AI to elicit empathy — a practice that would heighten anthropomorphism risks and ethical misunderstandings. Empathy in our model should be grounded in truthful interactions and human-centered values. Deliberately designing exaggerated or deceptive emotional expressions (e.g., a drone "screaming" in distress) does not foster genuine empathy but manipulates emotional response, potentially leading to moral disengagement or aversion. The goal of AI ethics is to promote human welfare — not to evoke unwarranted sympathy for machines. Healthy empathy mechanisms must be transparent and correctly targeted. For example, a search-and-rescue robot may use an urgent human-like tone to attract the attention of survivors — appealing to their hope for rescue, not seeking pity for the machine. In summary, we must design empathy-eliciting contexts wisely, acknowledging that AI has no emotions and cannot empathize, while leveraging its embodied presence to activate human moral emotions and direct them toward those who truly warrant care and responsibility.

In summary, the fact that "AI cannot empathize" does not weaken our theoretical framework — it instead underscores its necessity. Precisely because AI lacks emotion and moral sensitivity, we must emphasize the indispensable role of human empathy and ethical agency. Within our model, empathy functions both as a mechanism (eliciting and transmitting responsibility) and as a boundary (reminding us that AI remains a tool, never a source of moral feeling). It ensures that responsibility is ultimately borne by feeling and reasoning subjects — human beings.

The Western philosophical tradition offers two influential perspectives: Kant located morality in rational autonomy, while Hume rooted it in sympathy. AI, however, meets the conditions for neither. Our position may thus be viewed as integrative: both reason and empathy originate in humans, yet AI can serve as a unique touchstone for triggering and examining these capacities. Through empathy elicited in contexts involving AI, we reaffirm the irreplaceability of human moral subjectivity and clarify the central role of humanity within AI ethics and governance.

3. CLARIFYING THE HUMAN SUBJECTIVITY OF RESPONSIBILITY ATTRIBUTION

In summary, the fact that "AI cannot empathize" does not weaken our theoretical framework — it instead underscores its necessity. Precisely because AI lacks emotion and moral sensitivity, we must emphasize the indispensable role of human empathy and ethical agency. Within our model, empathy functions both as a mechanism (eliciting and transmitting responsibility) and as a boundary (reminding us that AI remains a tool, never a source of moral feeling). It ensures that responsibility is ultimately borne by feeling and reasoning subjects — human beings.

The Western philosophical tradition offers two influential perspectives: Kant located morality in rational autonomy, while Hume rooted it in sympathy. AI, however, meets the conditions for neither. Our position may thus be viewed as integrative: both reason and empathy originate in humans, yet AI can serve as a unique touchstone for triggering and examining these capacities. Through empathy elicited in contexts involving AI, we reaffirm the irreplaceability of human moral subjectivity and clarify the central role of humanity within AI ethics and governance (Baum et al., 2022). Both our model and the RRI framework are designed to ensure that this process of attribution is both coherent and legitimate: through institutional and empathetic mediation, responsibility for any AI behavior is ultimately mapped back onto human actors. This approach serves as a critique of anthropomorphic excess while simultaneously addressing concerns about "responsibility gaps." It reaffirms that although AI may act with apparent autonomy, accountability remains a distinctly human capacity and obligation.

Naturally, the "human" in this context is not limited to individuals but may encompass collective entities — teams, organizations, or even society as a whole forming a community of responsibility. As AI systems grow in complexity, "distributed responsibility" will become the norm, necessitating legal and ethical mechanisms that facilitate shared and coordinated

accountability. Yet, regardless of how responsibility is distributed, AI itself remains excluded from the moral circle. This demarcation is crucial to prevent two types of failures: first, anthropomorphic missteps such as punishing or empowering AI as if it were a moral agent — an approach that fails to correct human behavior (e.g., "deactivating" a faulty autonomous vehicle without holding the manufacturer accountable teaches nothing and achieves little); second, relinquishing human moral judgment to AI systems (e.g., allowing algorithms to allocate medical resources without human oversight or accountability). Only by insisting on human subjectivity in responsibility can we maintain ethical sovereignty over technology. Herein lies the tension within "non-subjective ethics": while it decentralizes agency at the operational level, it recenters humanity as the ultimate moral subject.

Finally, we acknowledge that this framework may have its limits. Should genuinely strong AI — with autonomous consciousness and emotion — emerge in the future, it might eventually cross the threshold into moral personhood. However, current evidence suggests that such a scenario remains distant. In this transitional era, it is imperative to clearly define responsibility: neither overestimating AI's moral capacity nor evading human obligations. We are dealing with an "intelligent Other," not an "intelligent subject" — an entity that behaves increasingly like a subject while remaining devoid of moral agency. This demands careful development of mechanisms, as described above, to prevent ethical ambiguity or vacuums. Even if AI were to attain moral personhood someday, it should occur only after robust safeguards are in place to prevent systemic accountability failures. In other words, AI governance must be firmly rooted in existing human ethical frameworks — not left to the hope that AI will autonomously develop ethical competence to resolve dilemmas we fail to address.

CONCLUSION

The emergence of embodied artificial intelligence presents a profound challenge to traditional ethical paradigms: how can we sustain established principles of responsibility when "non-human agents" participate in our social interactions? Through interdisciplinary theoretical inquiry, this paper proposes a potential pathway via a "non-subjective ethics." Central to this framework is the conception of responsibility not as an intrinsic attribute of a subject but as a relational phenomenon that manifests within interaction. Although embodied AI lacks subjectivity, its physical presence and behavioral expressions evoke moral responses in humans. From a phenomenological perspective, we trace how responsibility gradually emerges through three

stages — perception, situatedness, and appeal — within human-AI relations, ultimately anchoring itself in human moral agents.

This generative mechanism offers a novel approach to the problem of responsibility attribution: responsibility resides not within AI itself, but within the ethical concern awakened in humans through interaction with AI.

We further elevate this concept to the institutional level by reinterpreting the dimensions of RRI (Responsible Research and Innovation), demonstrating that governance principles such as anticipation, reflexivity, inclusion, and responsiveness collectively constitute a structure for the manifestation and implementation of responsibility. This enables responsibility to be integrated proactively, internalized, diversified, and operationalized throughout technological innovation.

Consequently, the future of embodied AI governance lies not in constructing AI as a moral agent, but in shaping human practices around AI to cultivate inherent ethical sensitivity and self-correcting capacity. This signifies a paradigm shift in technology governance: from an ethics of the subject (which requires individual actors to be morally self-sufficient) to an ethics of relations (which requires morality to be embodied in systemic interactions).

This paper also clarifies essential boundary conditions within this framework. We critique excessive anthropomorphism for potentially distorting and diluting responsibility and emphasize that AI possesses neither emotion nor empathy — hence, responsibility must be understood and borne exclusively by humans. These arguments consistently affirm one conclusion: the source and end of responsibility remains humanity itself, while AI serves as an impetus to reengage with this perennial moral truth in new ways.

Looking ahead, as embodied AI becomes more pervasive and sophisticated, the proposed non-subjective ethical framework must undergo empirical validation and further development. On one hand, as human-AI relationships intensify, finer-grained analyses of responsibility generation in real-time interactions — such as how dynamic trust and affective reactions shape responsibility judgments — will be needed. This may require integrating empirical insights from cognitive science and sociology to enrich the model.

On the other hand, public policy and legal systems must explore ways to institutionalize this "structure of responsibility manifestation," for instance, by establishing norms that compel AI developers to adhere to RRI dimensions or embedding ethical review into technical standardization processes. Furthermore, cultural variations in attitudes toward AI and ethical norms will influence how the "appeal of the Other" is perceived and how responsibility is conceptualized. While Western philosophy (e.g., Husserl,

Levinas) provides robust theoretical tools, Eastern relational ethics — such as Confucian and Daoist thought — may also offer valuable insights. This suggests a promising direction for future research: exploring cross-cultural perspectives on AI ethics to develop a more universal theoretical framework.

Regardless of technological evolution, one principle must remain unequivocal: humanity must not relinquish its sovereignty over ethics. The governance of embodied AI ultimately tests our own moral wisdom and courage. A non-subjective ethics does not diminish human significance — on the contrary, it underscores that humans are more indispensable than ever in the ethical domain, for we are responsible not only for our own actions but also for the behaviors of the "Other" we have created.

When we gaze into the mirror of artificial intelligence, it is in fact the Other who gazes back at us. Only by confronting this pressure and appeal of being seen can humanity continue to uphold its role as moral agents in the age of AI and shape a future where both technology and society evolve toward the good. As Kant proclaimed, humans are ends in themselves, never mere means; and Levinas further reminds us that the Other is the very origin of ends. Facing the future of embodied AI, we must uphold such philosophical clarity: even if using AI to govern AI presents a potential mechanism for technological governance, the underlying ethical responsibility can never — and should never — be transferred to AI. Let AI better serve the ends of humanity, yet always remember: the radiance of humanity begins with our responsible response to the Other.

REFERENCES

Asada, M. 2018. "Artificial Empathy." In *Diversity in Harmony : Insights from Psychology*, ed. by K. Shigemasu, S. Kuwano, T. Sato, and T. Matsuzawa, 19–41. Hoboken (NJ): John Wiley & Sons.

Baum, K., S. Mantel, E. Schmidt, and T. Speith. 2022. "From Responsibility to Reason-Giving Explainable Artificial Intelligence." *Philosophy & Technology* 35.

Burget, M., E. Bardone, and M. Pedaste. 2016. "Dimensions of Responsible Research and Innovation." In *INTED2016 Proceedings: 10th International Technology, Education and Development Conference*, ed. by L. Gómez Chova, A. López Martínez, and I. Candel Torres, 1008–1013. Valencia: IATED Academy.

Coeckelbergh, M. 2020. "Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability." *Science and Engineering Ethics* 26:2051–2068.

Gennaro, I. de, and R. Lüfter. 2024. "The Age of the Systemic Imperative. A Phenomenological Diagnosis of Social Responsibility." *HORIZON. Studies in Phenomenology* 13 (2): 587–609.

Georgia State University News Hub. 2024. "Study: Humans Rate Artificial Intelligence as More 'Moral' Than Other People." Georgia State University News Hub. Accessed June 1, 2025. https://news.gsu.edu/2024/05/06/study-humans-rate-artificial-intelligence-as-more-moral-than-other-people/.

Herzog, C. 2021. "Three Risks That Caution Against a Premature Implementation of Artificial Moral Agents for Practical and Economical Use." *Science and Engineering Ethics* 27 (1).

Husserl, E. 1970. *The Crisis of European Sciences and Transcendental Phenomenology [Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie]*. Trans. from the German by D. Carr. Evanston: Northwestern University Press.

———. 1983. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy [Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie]: First Book: General Introduction to a Pure Phenomenology [I. Buch: Allgemeine Einführung in die reine Phänomenologie]*. Trans. from the German by F. Kersten. Dordrecht and The Hague: Martinus Nijhoff Publishers.

Kawai, Y., T. Miyake, J. Park, et al. 2023. "Anthropomorphism-Based Causal and Responsibility Attributions to Robots." *Scientific Reports* 13:12234.

Lau, K. 2004. "Intersubjectivity and Phenomenology of the Other: Merleau-Ponty's Contribution." In *Space, Time and Culture*, ed. by D. Carr and Ch. Chan–Fai, 135–158. Contributions to Phenomenology 51. Dordrecht: Kluwer Academic Publishers.

Macnaghten, P., R. Owen, and R. Jackson. 2016. "Synthetic Biology and the Prospects for Responsible Innovation." *Essays in Biochemistry* 60 (4): 347–355.

Matthias, A. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183.

Montemayor, C., J. Halpern, and A. Fairweather. 2022. "In Principle Obstacles for Empathic AI: Why We Can't Replace Human Empathy in Healthcare." *AI & Society* 37 (4): 1353–1359.

Moor, J. H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems* 21 (4): 18–21.

Stahl, B. C. 2013. "Responsible Research and Innovation: The Role of Privacy in an Emerging Framework." *Science and Public Policy* 40 (6): 708–716.

Stilgoe, J., R. Owen, and P. Macnaghten. 2013. "Developing a Framework for Responsible Innovation." *Research Policy* 42 (9): 1568–1580.

Winfield, A., K. Michael, J. Pitt, and V. Evers. 2019. "Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems." *Proceedings of the IEEE* 107 (3): 509–517.

Zhu, G. 2006. "How Is Ethics as First Philosophy Possible? — On Levinas' Ethics and the Critique of the Violence of Being." *Journal of Nanjing University (Philosophy, Humanities and Social Sciences)* 43:24–32.

Cheng P., Zhang Zh. [Чэн П., Чжан Чж.*]* The Mechanism of Responsibility Generation and the Logic of Ethical Governance in Embodied Artificial Intelligence [Механизм формирования ответственности и логика этического управления во воплощенном искусственном интеллекте] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 123–151.

Пэн Чэн
к. филос. н., младший научный сотрудник, Китайский научно-исследовательский институт популяризации науки (Пекин); ORCID: 0009–0002–8474–4711

Чжихуэй Чжан
к. филос. н., профессор, Институт истории естественных наук Китайской академии наук (Пекин); ORCID: 0000–0003–1876–9312

Механизм формирования ответственности и логика этического управления в воплощенном искусственном интеллекте

Получено: 13.09.2025. **Рецензировано:** 30.09.2025. **Принято:** 18.10.2025.

Аннотация: Статья посвящена проблеме «разрыва ответственности», возникающего в связи с интеграцией воплощенного искусственного интеллекта в сферу социальных взаимодействий. Отказываясь от функционалистских моделей, которые приравнивают агентность ИИ к моральному статусу личности, авторы принимают феноменологическую перспективу и переосмысливают ответственность как феномен, «проявляющийся» в отношениях. Предлагается трехстадийная модель возникновения воплощенной ответственности: перцептивная презентация, ситуативная включенность, этический призыв. На этой основе проводится философская реконструкция четырех измерений ответственных исследований и инноваций (ООИ) — антиципации, рефлексивности, инклюзивности, реагирования, которые трактуются как «структура проявления ответственности». Статья предостерегает от этических рисков антропоморфизации, уточняет принципиальную неспособность ИИ к эмпатии и моральной агентности и утверждает, что ответственность в конечном счете должна оставаться на стороне человека и быть институционализирована через соответствующие механизмы управления.

Ключевые слова: воплощенный искусственный интеллект, атрибуция ответственности, этическое управление, феноменология ответственности, ответственные исследования и инновации.

DOI: 10.17323/2587-8719-2025-4-123-151.

Karpova, E. A. 2025. "Algorithmic Authority and Moral Responsibility : Rethinking Agency in the Age of Artificial Intelligence" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 152–166.

Elizaveta Karpova*

Algorithmic Authority and Moral Responsibility**

Rethinking Agency in the Age of Artificial Intelligence

Submitted: Sept. 15, 2025. **Reviewed:** Oct. 10, 2025. **Accepted:** Oct. 18, 2025.

Abstract: As artificial intelligence systems increasingly mediate decisions in domains such as healthcare, law, finance, and national security, traditional notions of moral agency and responsibility are being subjected to unprecedented scrutiny. Decisions once regarded as the sole prerogative of human judgment are now frequently delegated to or shaped by algorithmic processes, raising fundamental questions about the status of human agency in technologically mediated contexts. This article investigates the philosophical implications of what may be called algorithmic authority — the expanding normative power exercised by algorithm-driven systems over social, political, and ethical life. The rise of algorithmic authority destabilizes conventional frameworks of responsibility that presuppose a clear locus of agency in individual actors. When outcomes emerge from complex interactions between human intentions, institutional structures, and machine learning models, the boundaries of accountability become blurred. To address this challenge, the article argues for a framework of distributed moral responsibility, which better captures the hybrid and networked character of contemporary human-machine decision-making. Drawing on contemporary theories of agency, socio-technical systems, and ethics, this framework emphasizes that responsibility is not eroded but rather reconfigured: it becomes dispersed across multiple nodes, including designers, users, institutions, and the algorithms themselves as mediating agents. Ultimately, the article seeks to reconceptualize moral responsibility in a way that not only clarifies the ethical stakes of artificial intelligence but also provides guidance for developing normative principles suited to an algorithmically mediated world.

Keywords: Artificial Intelligence (AI), Medical Ethics, Moral Responsibility, Algorithmic Decision-Making, Human-AI Interaction, Ethical Expertise, Accountability.

DOI: 10.17323/2587-8719-2025-4-152-166.

INTRODUCTION

In recent years, artificial intelligence (AI) has moved from the periphery of technological imagination to the very center of decision-making processes that shape individual lives and collective destinies. From medical diagnostics

*Elizaveta Karpova, PhD Student in Philosophy; Research Assistant at HSE University (Moscow, Russia), ea.karpova@hse.ru, ORCID: 0009–0005–0499–7930.

to criminal sentencing algorithms, from predictive policing to automated loan approvals, AI systems are no longer passive tools; they increasingly function as agents of judgment, recommendation, and even command. This transformation brings into sharp relief a profound philosophical question: *Who — or what — is responsible when an algorithm makes a mistake?*

At the heart of this question lies a deeper conceptual challenge. Classical ethical theories have long been predicated on the assumption that moral agency resides in autonomous, rational human individuals. Responsibility, in this framework, is grounded in intention, consciousness, and the ability to deliberate. Yet AI systems, particularly those based on machine learning, operate without intention or consciousness. They are trained on data, optimized for patterns, and deployed within opaque infrastructures of code, institutions, and regulation. As a result, traditional models of ethical accountability often falter when applied to AI-driven contexts.

This article contends that the rise of algorithmic authority demands a fundamental rethinking of moral agency and responsibility. Instead of attempting to fit AI into traditional ethical frameworks, we must reconsider the very categories through which we evaluate responsibility. The paper begins by tracing the emergence of algorithmic authority as a new form of normative power. It then examines competing philosophical accounts of agency — both human and non-human—and explores the growing concern over *responsibility gaps* in automated systems. Finally, the article proposes a model of *distributed moral responsibility*, one that reflects the complex, layered, and relational structure of decision-making in the age of AI.

The task is urgent. As AI continues to permeate institutions and reshape practices, ethics must not remain reactive or reductive. Instead, it must evolve — conceptually and institutionally — to meet the challenges of an era in which moral choices are increasingly mediated by machines.

1. THE RISE OF ALGORITHMIC AUTHORITY

The notion of *authority* has traditionally been associated with persons or institutions recognized as legitimate sources of guidance, judgment, or command. From the sovereign in political theory to the physician in medical ethics, authority implies a relationship of trust, epistemic privilege, and normative force. In recent years, however, we have witnessed the emergence of a new, less tangible form of authority — *algorithmic authority* — which demands critical philosophical scrutiny.

Coined and developed in media and information studies (notably by Clay Shirky and later explored by Luciano Floridi), the term *algorithmic authority* captures a distinctive kind of power: one that derives not from human expertise or institutional legitimacy, but from computational processes themselves (Floridi, 2019; Shirky, 2008). This authority is embedded in systems that claim to produce reliable outputs — recommendations, classifications, decisions — by virtue of their algorithmic design and performance. In many cases, these outputs are treated as objective, neutral, or even superior to human judgment, thus acquiring de facto normative status.

1.1. AUTHORITY WITHOUT A FACE

Unlike traditional authorities, algorithms are faceless and impersonal. Their authority does not stem from charisma, reputation, or moral standing (Zerilli et al., 2019: 559). Rather, it is conferred by their perceived efficiency, data-driven accuracy, and capacity to scale across contexts. A diagnostic AI system, for instance, may outperform human radiologists in identifying certain types of tumors.[1] As a result, its recommendations may come to override or heavily influence clinical judgments — especially when institutional protocols are aligned with algorithmic outputs.

This shift is not merely technological; it is epistemological and moral. It affects how knowledge is produced, validated, and acted upon. It also transforms how responsibility is allocated. When a judge relies on a risk-assessment algorithm like COMPAS to determine bail or sentencing, who is ultimately responsible for the decision: the judge, the developers, the institution, or the algorithm itself? (Machine Bias, 2016) The very diffusion of authority leads to a diffusion — and often an erosion — of accountability.

1.2. PRACTICAL EXAMPLES AND DOMAINS OF CONCERN

The expanding domain of algorithmic authority is particularly evident in the following sectors.

- *Healthcare*: Clinical decision-support systems, such as IBM Watson for Oncology (now discontinued, but philosophically illustrative), once offered treatment recommendations based on large-scale data analysis. Physicians often deferred to these systems, even when human judgment might have raised doubts.

[1] See, for example, Esteva et al., 2017.

- ◇ *Law and Criminal Justice*: Algorithms used for predictive policing, parole recommendations, or sentencing guidance raise urgent questions about fairness, bias, and transparency. The opacity of these systems — often protected as proprietary — exacerbates public mistrust (Burrell, 2016).
- ◇ *Finance and Employment*: Credit-scoring algorithms and automated résumé filters determine access to loans and jobs. Here, algorithmic decisions may replicate or magnify existing social inequalities while eluding direct legal responsibility.
- ◇ *Warfare*: Autonomous weapons systems introduce the most extreme version of algorithmic authority: machines that may make life-or-death decisions with minimal or no human oversight (Sparrow, 2007: 72–74).

Each of these examples reflects a growing trend: the delegation of morally significant decisions to algorithmic systems whose internal workings are often inscrutable, even to their creators.

1.3. THE PHILOSOPHICAL STAKES

What distinguishes algorithmic authority from earlier technological systems is its *normative role* (The Ethics of Algorithms..., 2016: 65). These systems do not merely inform human judgment; they shape and sometimes replace it. They alter institutional practices and social expectations, often reinforcing the belief that machine-generated decisions are more reliable, unbiased, or objective than human ones. Yet this belief rests on shaky epistemic and moral ground. Algorithms reflect the assumptions, values, and limitations of their training data, their design parameters, and the social systems in which they are deployed.

Thus, algorithmic authority is not neutral (Eubanks, 2018: 139). It is a form of *constructed legitimacy* — one that bypasses traditional channels of ethical deliberation. It demands a philosophical response, not only in the form of critique but also in the development of new conceptual tools to address the ethical challenges it poses.

2. RETHINKING AGENCY: HUMAN, MACHINE, HYBRID

The concept of *agency* has long occupied a central place in ethical theory, grounded in notions of autonomy, intentionality, and moral responsibility (Korsgaard, 1996: 7). In Kantian and post-Kantian traditions, agency is fundamentally human: to act is to will, to deliberate, to choose. Yet in the context of artificial intelligence, such assumptions are increasingly

strained. As AI systems participate in decisions with ethical consequences — often without direct human oversight — we are compelled to revisit and reconsider our understanding of what it means to be an agent.

2.1. CLASSICAL NOTIONS OF AGENCY AND THEIR LIMITS

In traditional moral philosophy, agency is typically associated with rational deliberation and moral accountability. The agent is someone who can form intentions, understand norms, and be held responsible for their actions. This model is anthropocentric and deeply embedded in legal and ethical practices. However, AI systems — especially those based on machine learning — do not operate on intention or moral deliberation. They process data, optimize outputs, and "learn" correlations. As such, they lack core features of classical agency, including self-reflection, moral reasoning, and accountability. To call them *agents* in the classical sense would be a category mistake (Searle, 1980: 431).

Yet AI systems increasingly act *as if* they were agents: they interact with humans, make autonomous recommendations, and adapt to new environments. Their behavior has consequences indistinguishable from intentional action in practical terms, even if philosophically they lack intention. This raises the question: *Can we develop a more nuanced concept of agency that accommodates these new actors without falling into anthropomorphism or ethical confusion?*

2.2. FROM ARTIFICIAL AGENTS TO DISTRIBUTED AGENCY

A growing body of work in philosophy of technology and science and technology studies (STS) proposes a shift away from individualistic models of agency (Latour, 2005: 8). Instead, it advocates for a view of *distributed agency*, in which actions emerge from networks of human and non-human actors. On this view, agency is not a substance or property but a relational effect: it is enacted through interaction, coordination, and infrastructure.

This perspective resonates with theories such as:

- Actor-Network Theory (ANT), which treats both humans and non-humans as actants in social assemblages;
- Extended mind theories, which locate cognition (and agency) across the brain, body, and environment;
- Postphenomenology, which emphasizes the mediating role of technology in human perception and action.

Within such frameworks, AI systems do not *possess* agency in the classical sense, but they *participate* in agential configurations (Verbeek, 2011). An autonomous vehicle, for instance, acts within a complex ecology of sensors, algorithms, regulations, urban infrastructure, and human supervision. Responsibility, accordingly, is not located in a single node, but distributed across the system.

2.3. HYBRID MORAL AGENTS: BETWEEN AUTONOMY AND DELEGATION

Some theorists have suggested that the notion of *hybrid agency* may offer a useful middle ground (Coeckelbergh & Calo, 2015: 529). Hybrid agents are composite systems — part human, part machine — in which decision-making unfolds through a dynamic interplay. In these cases, human agents retain partial control or oversight, but their actions are shaped and constrained by algorithmic mediation. Consider, for example, the use of clinical decision support systems (CDSS) in hospitals.[2] A physician may remain the formal decision-maker, yet their judgment is shaped by algorithmic recommendations, interface design, legal liability, and time pressure. Here, the "agent" is neither the doctor nor the AI alone, but the assemblage that links them. Ethical responsibility, likewise, must be rethought in terms of this hybridity.

By reconfiguring our models of agency, we can move beyond the sterile binary of *human versus machine* and begin to articulate ethical frameworks that better reflect the socio-technical reality of contemporary decision-making.

3. RESPONSIBILITY GAPS AND THE ETHICS OF DELEGATION

As algorithmic systems increasingly operate in high-stakes environments — autonomous vehicles, predictive policing, clinical diagnostics — the traditional frameworks of moral and legal responsibility begin to falter. When things go wrong, it is often unclear who should be held accountable: the developer, the deploying institution, the end-user, or the system itself? This ambiguity has given rise to what scholars term "responsibility gaps," structural voids in accountability that emerge when outcomes are shaped by systems that resist full human control or comprehension.

[2]See, for example, Annas, 2012; Jotterand & Bosco, 2021.

3.1. THE EMERGENCE OF RESPONSIBILITY GAPS

Philosopher Andreas Matthias coined the term "responsibility gap" in the context of autonomous weapons systems — technologies capable of lethal action without direct human command (Matthias, 2004: 176). The challenge, he argued, lies in the fact that these systems may act unpredictably due to their learning-based architectures. Traditional attribution models (based on intent or foreseeability) no longer apply cleanly when the behavior of the agent cannot be traced back to a human actor with sufficient knowledge or control. The problem is not confined to military contexts. Similar gaps arise in algorithmic trading, healthcare diagnostics, and criminal justice (Danaher, 2016: 250–251). When an AI-based risk assessment tool recommends a higher sentence based on biased data, it may be difficult to identify a single culpable party — especially when the model is opaque, proprietary, and complex.

Responsibility, under such conditions, is neither absent nor irrelevant — it is displaced, dispersed, and distorted. Ethics must account for these displacements not by collapsing the issue into a nihilistic "no one to blame" stance, but by rethinking the very architecture of delegation and moral liability.

3.2. DELEGATED AGENCY AND THE PROBLEM OF CONTROL

Delegation is a pervasive feature of social and institutional life (Nyholm, 2018: 1211). We delegate tasks to subordinates, institutions, and tools. What makes delegation ethically permissible is that the delegator retains *control*, *oversight*, and *accountability* for the outcome. When machines act in ways that defy their designer's or the user's expectations, that triad is broken. Control becomes probabilistic, oversight becomes partial, and accountability becomes elusive.

One response to this challenge is to treat algorithmic systems as *moral proxies* — tools that act on behalf of humans within specified constraints. But proxies can fail. They can misrepresent the values of those they stand in for or act in unanticipated ways (Coeckelbergh, 2010: 66). The analogy to human delegation begins to unravel when proxies become adaptive, opaque and non-transparent.

As a result, some scholars have argued for the need to develop *new models of responsibility* that acknowledge this partiality. These include *forward-looking responsibility* (focused on improving systems and reducing harm) and *distributed responsibility* (allocating accountability across networks of actors and designers) (Van de Poel & Sand, 2021: 4773–4774). However, such

models raise difficult questions: How do we ensure justice for victims? Who compensates for harms? Can diffuse responsibility still retain moral weight?

3.3. ETHICAL DESIGN AND INSTITUTIONAL ACCOUNTABILITY

To address responsibility gaps, it is not enough to seek new individual scapegoats; the solution must be structural. One promising direction lies in what is often called *ethical design*: embedding ethical considerations into the very architecture of AI systems (AI4People..., 2018: 701). This includes transparency, explainability, auditability, and human-in-the-loop mechanisms. Yet ethical design must be matched by *institutional responsibility*. Organizations that develop or deploy AI must assume proactive roles: conducting ethical impact assessments, establishing redress mechanisms, and ensuring that their delegation to machines is not a form of moral outsourcing (Wagner, 2019).

In this context, ethics becomes not a post-hoc response to harm, but a precondition of technological legitimacy. It asks not only *who is responsible after the fact*, but *how responsibility is structured and shared in advance*. Bridging the responsibility gap thus requires not simply attribution, but design — ethical, institutional, and philosophical.

4. TOWARD A FRAMEWORK OF DISTRIBUTED MORAL RESPONSIBILITY

The emergence of intelligent systems capable of autonomous decision-making has exposed a fundamental tension in ethical theory and practice: the inadequacy of traditional, individual-centered models of moral responsibility. When actions and outcomes are co-produced by a heterogeneous network of human and non-human agents — engineers, algorithms, platforms, users, institutions — assigning moral liability to a single source becomes both philosophically and practically untenable. This phenomenon, often framed as the "responsibility gap," calls for a reconceptualization of how moral responsibility is understood and allocated within complex socio-technical systems (Matthias, 2004: 179–180).

In this context, we propose a shift toward *distributed moral responsibility* — a framework grounded in relational, process-oriented, and multi-actor perspectives that reflect the hybrid nature of human-machine interaction. Rather than seeking a singular locus of accountability, this approach emphasizes shared, overlapping, and context-sensitive forms of responsibility that correspond to varying degrees of influence, foresight, and agency within the system.

4.1. DISTRIBUTING RESPONSIBILITY ACROSS ACTORS

Distributed moral responsibility begins by recognizing that moral agency is not confined to isolated individuals but emerges through interactions within structured environments. In algorithmic ecosystems, multiple agents — human and artificial — participate in the generation of outcomes. These include:

- Designers and developers, who embed ethical assumptions into models and code architectures;
- Deployers, such as corporations or institutions, who configure and implement systems in real-world settings;
- End-users, who interact with and may be guided or constrained by algorithmic outputs;
- Regulators and policymakers, who shape the institutional and legal frameworks in which these technologies operate.

Each of these actors operates within different spheres of control and epistemic access. For instance, developers may understand system architecture but lack insight into its downstream applications, while regulators may have oversight power without the technical granularity. A model of distributed responsibility must therefore correlate responsibility with actual and potential capacities for action, including the ability to anticipate risks, intervene meaningfully, and reflect on outcomes (Gunkel, 2012: 143).

Moreover, while artificial agents cannot be said to possess moral agency in the full sense — given their lack of consciousness, intentionality, and capacity for moral reasoning — their actions can still mediate or amplify human intentions. In this light, machines become moral intermediaries, requiring that their integration into decision-making processes be accompanied by new modes of ethical oversight and co-responsibility.

Importantly, this distribution is not meant to dilute or deflect responsibility, but rather to map it more accurately onto the networked structure of action and causality. Recognizing distributed responsibility allows us to avoid both the "scapegoating" of frontline users and the abdication of accountability by upstream actors.

4.2. DIMENSIONS OF RESPONSIBILITY: FORWARD- AND BACKWARD-LOOKING

An adequate framework must also differentiate between two key dimensions of responsibility:

- *Forward-looking responsibility*, which emphasizes proactive duties such as the prevention of harm, the design of accountable systems, and the establishment of meaningful human oversight;

- *Backward-looking responsibility*, which focuses on determining liability after an adverse event or ethical failure, including attribution, compensation, and institutional learning.

Both dimensions are indispensable. Forward-looking responsibility fosters ethical anticipation and precaution, crucial in the design phase of AI systems. This includes practices such as ethical impact assessments, participatory design, and scenario planning. In contrast, backward-looking responsibility ensures that harms are acknowledged and addressed, thus maintaining public trust and reinforcing the legitimacy of technological governance.

Central to both is the idea of "meaningful human control" — a normative standard according to which human actors must remain sufficiently involved in and accountable for the actions of autonomous systems (Santoni de Sio & van den Hoven, 2018: 2). This principle ensures that responsibility remains traceable and that moral reflection is not bypassed in favor of purely instrumental efficiency.

4.3. EMBEDDING RESPONSIBILITY INTO SYSTEMIC DESIGN AND GOVERNANCE

To operationalize distributed responsibility, we must move beyond abstract principles and embed ethical safeguards at multiple levels of design and governance. This involves:

- *Transparency and explainability*: Making algorithmic processes intelligible to relevant stakeholders, including developers, users, and regulators. Interpretability is not only a technical challenge but a moral imperative — it enables accountability and informed consent (Doshi-Velez & Kim, 2017).
- *Human-in-the-loop and human-on-the-loop mechanisms*: Preserving the ability of humans to intervene, override, or guide autonomous systems, especially in high-stakes domains such as healthcare, policing, or finance.
- *Ethical oversight infrastructures*: Establishing institutional mechanisms such as ethics boards, algorithmic audit trails, and redress systems that can respond to ethical concerns post-deployment (Mittelstadt, 2019: 501).
- *Responsibility mapping*: Creating tools to visualize and track responsibility across the algorithmic supply chain — from data collection to model training, deployment, and use (Amoore, 2020: 89). This mapping makes visible the roles and responsibilities that are often obscured by technical complexity.

4.4. RESISTING THE TEMPTATION OF MORAL OUTSOURCING

Finally, we must confront a pervasive temptation in contemporary techno-ethics: the outsourcing of moral judgment to machines. Delegating decisions to algorithmic systems may offer efficiency or consistency, but it also risks a form of moral disengagement (Coeckelbergh & Calo, 2015: 531). When humans defer to automated outputs uncritically, they may abdicate their ethical responsibilities and undermine the very possibility of accountability.

A distributed framework resists this tendency by reaffirming the centrality of human moral agency — not as an isolated sovereign will, but as a situated, relational practice embedded in social and technological contexts. It invites us to cultivate new forms of ethical competence: interdisciplinary communication, reflexive design, and collective deliberation.

Ultimately, distributed moral responsibility is not only a response to technical complexity — it is a normative commitment to rethinking responsibility itself in an age of entangled agencies and algorithmic mediation.

5. CONCLUSION AND FUTURE DIRECTIONS

The growing integration of artificial intelligence into decision-making infrastructures presents a profound challenge to established paradigms of moral responsibility. Traditional models — anchored in individual intentionality, linear causality, and binary agency — are increasingly misaligned with the distributed, opaque, and hybrid character of socio-technical systems. As this paper has argued, meeting this challenge requires more than incremental ethical adjustments or after-the-fact accountability mechanisms. It demands a conceptual reframing of responsibility itself, grounded in philosophical reflection, institutional innovation, and technological design.

We have proposed the framework of *distributed moral responsibility* as a response to the epistemic and normative dislocations induced by algorithmic agency. This framework acknowledges that responsibility must be plural, situated, and dynamically allocated across a heterogeneous network of human and non-human actors. By foregrounding the roles of designers, deployers, regulators, and users — while retaining space for human moral judgment and collective reflexivity — it offers a structure for both proactive and retrospective ethical accountability. Crucially, it resists the temptation to dissolve responsibility into ambiguity or automation. Instead, it insists on tracing moral obligations along the lines of influence, control, and awareness.

Yet, this is only a starting point. Several pressing directions for future research and institutional development remain:

1. Recalibrating Legal and Ethical Norms. Legal frameworks around liability and responsibility are ill-equipped to accommodate systems that act autonomously, learn from data, and evolve over time (Floridi & Cowls, 2021). New regulatory architectures are needed — ones that can account for partial, shared, and forward-looking responsibility without collapsing into moral diffusion. Bridging the gap between ethical theory and legal practice will be a defining challenge of the next decade.

2. Designing for Responsibility. Ethical responsibility must be embedded not only in abstract principles but in the very architecture of intelligent systems (Santoni de Sio & van den Hoven, 2018). This calls for the further development of *responsibility-sensitive design practices*, including transparency-enhancing interfaces, traceability mechanisms, and participatory design methodologies. Technological design is not ethically neutral — it actively shapes what forms of action and reflection are possible (Verbeek, 2011).

3. Cultivating Ethical Agency in Human Actors. As we delegate more decisions to machines, we must also cultivate new capacities for human ethical agency: critical awareness, deliberative engagement, and institutional responsibility. Education in AI ethics should not be confined to engineers or philosophers — it must become a cross-sectoral and civic concern. Moral responsibility is not just about preventing harm, but about forming communities capable of sustained ethical reflection (Danaher, 2017).

4. Rethinking the Concept of Agency Itself. Finally, the rise of AI compels us to revisit the very notion of agency. If agency is no longer the exclusive domain of conscious, autonomous individuals, how should we reconceive it in relational, procedural, or systemic terms? What does it mean to act responsibly in a world where actions are co-produced by algorithms, infrastructures, and institutions? These questions require renewed dialogue between philosophy, sociology, cognitive science, and computer science.

In conclusion, the future of moral responsibility in the age of AI is not a matter of preserving old categories, but of rethinking them in light of technological transformations. Responsibility must remain a human concern — even, and especially, when it is shared across systems (Bryson, 2018). Our task is not to retreat from complexity, but to articulate new forms of moral understanding that are adequate to it.

REFERENCES

Amoore, L. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others.* Durham: Duke University Press.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. "Machine Bias." ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Annas, G. J. 2012. "Doctors, Patients, and Lawyers — Two Centuries of Health Law." *New England Journal of Medicine* 367:445–450.

Bryson, J. J. 2018. "Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20 (21): 15–26.

Burrell, J. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1).

Coeckelbergh, M. 2010. "Artificial Agents, Good Care, and Modernity: Towards a Technofuture-Oriented Ethics of Care." *Medicine, Health Care and Philosophy* 13 (1): 61–68.

Coeckelbergh, M., and R. Calo. 2015. "AI Ethics; Robotics and the Lessons of Cyberlaw." *California Law Review* 103 (3): 513–563.

Danaher, J. 2016. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29 (3): 245–268.

——— . 2017. "Will Life Be Worth Living in a World without Work? Technological Unemployment and the Meaning of Life." *Science and Engineering Ethics* 23 (1): 41–64.

Doshi-Velez, F., and B. Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv. https://arxiv.org/abs/1702.08608.

Esteva, A., B. Kuprel, R. A. Novoa, et al. 2017. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542:115–118.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Floridi, L. 2019. *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford University Press.

Floridi, L., and J. Cowls. 2021. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review*.

Floridi, L., J. Cowls, M. Beltrametti, et al. 2018. "AI4People— An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28:689–707.

Gunkel, D. J. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge (MA): MIT Press.

Jotterand, F., and C. Bosco. 2021. "Keeping the 'Human in the Loop' in the Age of Artificial Intelligence: Accountability and Values in Medical AI." *Journal of Medical Ethics* 47 (6): 389–393.

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge (MA): Cambridge University Press.

Latour, B. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Matthias, A. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6 (3): 175–183.

Mittelstadt, B. 2019. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1:501–507.

Mittelstadt, B., P. Allo, M. Taddeo, et al. 2016. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3 (2).

Nyholm, S. 2018. "Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24:1201–1219.

Santoni de Sio, F., and J. van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5.

Searle, J. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–457.

Shirky, C. 2008. *Here Comes Everybody: The Power of Organizing Without Organizations*. New York: Penguin Press.

Sparrow, R. 2007. "Killer Robots." *Journal of Applied Philosophy* 24 (1): 62–77.

Van de Poel, I., and M. Sand. 2021. "Varieties of Responsibility: Two Problems of Responsible Innovation." *Synthese* 198 (19): 4769–4787.

Verbeek, P.-P. 2011. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press.

Wagner, B. 2019. "Ethics as an Escape from Regulation: From Ethics-Washing to Ethics-Shopping?" In *Being Profiled. Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, ed. by E. Bayamlioglu, I. Baraliuc, L. Janssens, and M. Hildebrandt, 84–88. Amsterdam: Amsterdam University Press.

Zerilli, J., A. Knott, J. Maclaurin, and C. Gavaghan. 2019. "Algorithmic Decision–Making and the Control Problem." *Minds and Machines* 29 (4): 555–578.

Karpova E. A. [Карпова Е. А.] Algorithmic Authority and Moral Responsibility [Алгоритмическая власть и моральная ответственность] : Rethinking Agency in the Age of Artificial Intelligence [переосмысление агентности в эпоху искусственного интеллекта] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 152–166.

Елизавета Карпова
аспирантка, стажер-исследователь, НИУ ВШЭ (Москва); orcid: 0009–0005–0499–7930

Алгоритмическая власть и моральная ответственность

переосмысление агентности в эпоху искусственного интеллекта

Получено: 15.09.2025. **Рецензировано:** 10.10.2025. **Принято:** 18.10.2025.

Аннотация: По мере того как системы искусственного интеллекта все активнее участвуют в принятии решений в таких сферах, как здравоохранение, право, финансы и национальная безопасность, традиционные представления о моральном агентстве и ответственности оказываются под серьезным давлением. Решения, ранее принадлежавшие исключительно человеческому суждению, все чаще формируются под воздействием алгоритмов, что вызывает вопросы о статусе человеческой агентности в условиях технологически опосредованных практик. В статье рассматриваются философские последствия феномена алгоритмической власти — возрастающего нормативного влияния алгоритмических систем на социальную и этическую жизнь. Рост алгоритмической власти ставит под сомнение адекватность классических моделей ответственности, основанных на представлении о четко определенном субъекте. Когда результаты возникают из взаимодействия человеческих намерений, институциональных структур и алгоритмов машинного обучения, границы подотчетности размываются. В качестве альтернативы предлагается концепция распределенной моральной ответственности, отражающая сетевой и гибридный характер совместного принятия решений человеком и машиной. Опираясь на современные теории агентности, социотехнических систем и этики, статья утверждает, что ответственность не исчезает, а трансформируется: она распределяется между разработчиками, пользователями, институтами и алгоритмами как посредниками. Такой подход обеспечивает более адекватное понимание подотчетности и формирует нормативные ориентиры, необходимые в условиях алгоритмического управления.

Ключевые слова: искусственный интеллект (ИИ), медицинская этика, моральная ответственность, алгоритмическое принятие решений, взаимодействие человека и ИИ, этическая экспертиза, подотчетность.

DOI: 10.17323/2587-8719-2025-4-152-166.

Практическая философия

Studies: Practical Philosophy

Shishkov, A. V. 2025. "The Short History of Development of Object-Oriented Ontology" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 169–193.

Andrey Shishkov*

The Short History of Development of Object-Oriented Ontology**

Submitted: Aug. 28, 2025. **Reviewed:** Oct. 10, 2025. **Accepted:** Oct. 18, 2025.

Abstract: The article traces the history of the development of object-oriented ontology (OOO), a contemporary school of thought within post-continental philosophy, founded by Graham Harman in the late 1990s. It outlines OOO's emergence in the early works of Harman, its participation in speculative realism movement, its institutionalization as an independent philosophical direction and school of thought, and its expansion into different fields. The analysis is structured through a clear periodization, charting OOO's evolution from its inception as Harman's solitary project (1997–2006) to its participation in the birth of speculative realist movement and the subsequent formation of canonical OOO quartet alongside Levi Bryant, Ian Bogost, and Timothy Morton (2007–2011). The third period details OOO's institutionalization and prolific expansion beyond philosophy in ecology, art, architecture, archeology, religion, and other fields (2011–present). The article concludes by evaluating OOO's current state of maturity, acknowledging signs of internal diversification and theoretical reassessment among its founders, while asserting its enduring legacy as a distinct and influential school of thought that successfully challenged the anthropocentric paradigm of much preceding continental philosophy. The author makes a hypothesis about the beginning of fourth period.

Keywords: Graham Harman, Guerilla Metaphysics, Ian Bogost, Levi Bryant, Object-Oriented Ontology, Quadruple Object, Speculative Realism, Timothy Morton.

DOI: 10.17323/2587-8719-2025-4-169-193.

Object-oriented philosophy (OOP), or object-oriented ontology (OOO) as a direction, intellectual movement, and school of thought, emerged over than a quarter of a century ago. However, despite its fame and influence (especially in the fields of environmental criticism, art, architecture, design, etc.), there have been no attempts to write the history of its formation and make a historical and philosophical analysis of this phenomenon in the philosophical research literature.

Partly due to this, its importance is underestimated in world philosophy, although, in my opinion, its founder, Graham Harman (b. 1968), deserves to occupy his place in the history of philosophy among first-rank thinkers, such as R. Descartes, I. Kant, G. W. F. Hegel, F. Nietzsche, L. Wittgenstein,

*Andrey Shishkov, PhD Student in Philosophy at the Russian State University for the Humanities (Moscow, Russia), andrey.v.shishkov@gmail.com, ORCID: 0009–0001–4270–1900.

M. Heidegger, and others. In the article under consideration, I will try to fill this gap and place OOO in the context of contemporary post-continental philosophy, trace the connections with other philosophers, trends, and schools, highlight the main stages of development and propose its periodization.

Harman first used the name object-oriented philosophy in April 1999 in his lecture at Brunel University of London. Among other participants, the French philosopher and social theorist Bruno Latour (1947–2022) was present; he later became a kind of godfather of OOP and Harman's main intellectual sparring partner for many years. At that time, object-oriented philosophy was perceived as a less successful twin of actor-network theory (ANT) due to the similarity of its main ideas.[1]

Object-oriented philosophy began as Graham Harman's project, and for a long time he remained the only OO-philosopher. After the famous seminar at Goldsmiths College, University of London, which launched the speculative realist movement (2007), there was an explosion of interest in it. In 2008–2010, three more philosophers joined Harman's project, forming the canonical quartet of object-oriented philosophers: Timothy Morton (b. 1968), Levi Bryant (b. 1974), and Ian Bogost (b. 1976). In 2009, Bryant proposed rebranding the project by changing the word "philosophy" to "ontology"— this is how the acronym OOO, or triple O, appeared.

OBJECT-ORIENTED PHILOSOPHY IN THE CONTEXT OF POST-CONTINENTAL THOUGHT

Object-oriented ontology can be classified as a post-continental direction in philosophical thought. The prefix *post-* indicates a certain problematic character of this positioning — both within and outside the continental tradition. As John Mallarkey notes, the prefix *post-* means that the thinkers in this group "represent a real change in the intellectual current, one that both retains and abandons parts of what previously went under the rubric of 'Continental philosophy'" (Mallarkey, 2007: 1). Mallarkey identifies four French thinkers — Gilles Deleuze (1925–1995), Alain Badiou (b. 1937), Michel Henry (1922–2002), and François Laruelle (1937–2024) — as post-continental thinkers. However, Paul Ennis places the philosophers associated with OOO among the "post-continental voices" of contemporary philosophy

[1]Sometimes this has caused real confusion. For example, in the *Speculative Grace: Bruno Latour and Object-Oriented Theology* (Miller, 2013), the object-oriented approach is understood exclusively as the actor-network theory.

(Ennis, 2010), although in another work he characterizes them as "continental realism" (Ennis, 2011).

Indeed, Harman and Bryant, by virtue of their philosophical training, belong to the (post)continental tradition: the former studied French phenomenology and Martin Heidegger (1889–1976), the latter studied Gilles Deleuze and Jacques Lacan (1901–1981). Morton and Bogost came to philosophy from literary criticism, and they also turned to continental thought in their philosophical works. On the other hand, these philosophers seek to overcome the framework of continental thought, which they consider correlationist and anti-realist (Bryant et al., eds, 2011: 3–4). Harman even contrasts his approach with both the continental and analytic traditions (Harman, 2005: 1).

OOO has become part of several philosophical turns: speculative (Bryant et al., eds, 2011), ontological (Holbraad & Pedersen, 2017), and non-human turns (Grusin, 2015). And although the multitude of self-proclaimed "turns" has significantly diminished the significance of this approach to the history of philosophy, they can be interpreted as separate aspects of a change in the philosophical paradigm, drawing on Thomas Kuhn's concept. All these "turns" are united by a critique of the anthropocentrism of the preceding philosophical paradigm. If Immanuel Kant carried out a "Copernican revolution" in philosophy, then OOO, with its guerrilla metaphysics, can be seen as participating in an "anti-Copernican coup."

In their programmatic article "Towards a Speculative Philosophy," Levi Bryant, Nick Srnicek, and Graham Harman argue that "despite the vaunted anti-humanism," thinkers from such areas of continental philosophy as phenomenology, structuralism, post-structuralism, deconstruction, and postmodernism "give us is less a critique of humanity's place in the world, than a less sweeping critique of the self-enclosed Cartesian subject." At the same time, human being remains at the center of these approaches, and reality in them appears as a "correlate of human thought," which is why these areas of continental philosophy can be called anti-realistic. Thus, overcoming anthropocentrism begins only with speculative philosophy, to which OOO belongs (Bryant et al., eds, 2011: 2–3).

OBJECT-ORIENTED PHILOSOPHY BEGINS (1997–2006)

The founder of OOO, American philosopher Graham Harman, grew up in the Midwest, in the small town of Mount Vernon, Iowa. Little Graham's interest in philosophy was instilled by his mother, who enrolled the 13–14-year-old boy in philosophy classes. By the age of 16, this interest became

more conscious, and Harman felt that "at heart I am a metaphysician" (Harman & Pinho, 2020). In 1990, he received a Bachelor of Arts degree from St. John's College in Annapolis, Maryland and entered the Master of Philosophy program at the University of Pennsylvania (Philadelphia), where Harman's primary field was phenomenology. In 1991, he defended his master's thesis under the supervision of Prof. Alphonso Lingis (1933–2025), a distinguished American phenomenologist and translator of the works of Emmanuel Levinas (1906–1995) and Maurice Merleau-Ponty (1908–1961) into English. All three had a major influence on Harman's formation as a philosopher.

In 1996, Harman began his doctoral studies at DePaul University in Chicago. As early as 1991–1992, he had the idea that "the whole of Heidegger's philosophy can be understood from the tool-analysis[2] of *Being and Time*" (Harman, Morozov & Myshkin, 2015: 13),[3] and he wanted to demonstrate this in his doctoral research. However, until 1997, when his philosophical thinking changed, there was no hint of realist philosophy or an object-oriented approach in Harman's work. When he began his doctoral studies, he had no idea how to implement this project; he was at an impasse and made a living by reviewing sports events. Unlike Lingis, whom Harman always spoke warmly of, he was never able to find common ground with his new supervisor, the young professor William McNeill (b. 1961). The conservative Heideggerian, McNeill was not too welcoming of his unorthodox interpretation of Heidegger: "Even in later years he could not introduce me to an audience without making snarky remarks" (ibid.). Harman still looks back on his time as a doctoral student in rather dark terms.

In the summer of 1997, Harman encountered two books that freed him "from Heidegger's long shadow" (ibid.): *Process and Reality* (1929) by Alfred North Whitehead (1861–1947) and *On Essence* (1962) by the lesser-known Basque Catholic philosopher Javier Zubiri (1898–1983).

> Alfred North Whitehead's *Process and Reality* liberated me from the Kantian burden of Heidegger's thinking, where the relationship between a human person and the world is thought as more important than the relationship between two inanimate objects... After Zubiri, I became convinced that any entity must be construed in a radically non-relational sense (ibid.: 13–14).

[2]Harman uses the term *tools* to refer to Heidegger's *das Zeug*.

[3]Harman provides many biographical details of his early work in the preface to the Russian-language edition of *The Quadruple Object*.

In the autumn of that year, Harman made several attempts to put his ideas into writing. One such attempt was a lecture on the theory of objects in Heidegger and Whitehead, which he gave on Halloween evening to graduate students and several professors at his university (Harman, 2010: 36). The insight came at Christmas in 1997, when Harman was able to formulate the central tenet of object-oriented philosophy for the first time: "Far from being a coherent system, as Heidegger supposes, the world is partly connected by a chain of autonomous individual beings, each of which is partly hidden from the others" (Harman, Morozov & Myshkin, 2015: 14).

Later Harman explained how such philosophically opposed thinkers co-existed within his own approach:

> Retrospectively, the tension between the two authors is obvious. But I was more intrigued by how they complemented each other. Although Javier Zubiri's rejection of the relational approach clearly conflicted with Whitehead's metaphysics, it was Whitehead who liberated me from the anthropocentric defect in Heidegger's philosophy. The result of this combined influence was an early form of what is now known as object-oriented philosophy: a cosmological philosophy that deals with object-oriented relations, including humans as objects, and a philosophy in which objects are dark surpluses, never fully expressed in any relations (ibid.).

In addition to Whitehead and Zubiri, "the perfect medicine for my post-Heideggerian hangover" (Harman, 2010: 8) was reading Bruno Latour, whose works Harman first encountered in 1998. He entered into a correspondence with the French philosopher that marked the beginning of their friendship and intellectual competition.

On March 17, 1999, Harman successfully defended his doctoral thesis, entitled *Tool-Being: Elements of a Theory of Objects*. After slightly reworking the text of the dissertation, Harman published it as a monograph, *Tool-Being: Heidegger and the Metaphysics of Objects* (Harman, 2002). The first two chapters of the book constitute the very same unorthodox commentary on the tool-analysis and the main concepts of Heidegger's philosophy (*Dasein*, *Ereignis*, being, time, truth, language, technology, etc.), whose essence is to shift the focus from *Dasein* and to see in the tool-analysis of the German philosopher "an ontology of *objects themselves*" (ibid.: 1). In the third chapter, Harman offered the first detailed sketch of object-oriented philosophy, turning to the ideas of Whitehead, Zubiri and Levinas. Two central concepts of object-oriented philosophy appear in this work: *withdraw*, which places the object in a *private vacuum*, isolating it from all direct

relations, and *philosophy of human access*, which he later replaced with the term *correlationism* proposed by Quentin Meillassoux.

After defending his dissertation, Harman entered the academic job market and sent his résumé to several dozen educational institutions.[4] He received a response from only one — the American University in Cairo, Egypt, where he was offered the position of associate professor of philosophy. In addition to teaching and research, Harman also received an administrative position as vice provost for research. Cairo became his main place of work until 2016. In 2005, his second book, *Guerrilla Metaphysics: Phenomenology and the Workshop of Things* (Harman, 2005), was published, which Harman described as a sequel to *Tool-Being*. Initially, this book consisted of two different manuscripts: one was devoted to the problem of perceiving objects in the context of carnal phenomenology, to which Harman refers Levinas, Merleau-Ponty, and Lingis. One of the chapters in the first part also discusses the phenomenology of Dominique Janicaud (1937–2002) and the theory of intentional objects of Edmund Husserl (1859–1938), which would later play an important role in the formation of his own theory of objects.

The "guerrilla" character of Harman's metaphysics indicates that it stands in opposition to the traditional metaphysical mainstream, against which the main philosophical criticism of metaphysics as ontotheology is deployed (and Harman agrees with this criticism), At the same time, however, his metaphysics is able to respond to this criticism from an unexpected angle, winning back the right of metaphysics to exist. Guerrilla metaphysics simultaneously attacks traditional metaphysics as ontotheology, on the one hand, and nonmetaphysical phenomenology (Janicaud) and post-metaphysics (Marion, Caputo) on the other.

In the second part, Harman first proposed the concept of *vicarious causation*, which replaced the less successful concept of *occasional cause* from *Tool-Being*. Among the sources that influenced the formation of this concept, one should mention the theology of Arabic medieval occasionalism, which Harman became acquainted with in Egypt. Vicarious causation was supposed to explain how objects removed from all relations into private vacuums interact with each other. Here he also proposed the theory of metaphor, which is key to all subsequent object-oriented philosophy, based on the works of the Spanish philosopher José Ortega y Gasset (1883–1955).

[4]Harman told this story at the presentation of the second Russian-language edition of his book *Object-Oriented Ontology: A New "Theory of Everything"* at the Moscow branch of the Piotrovsky bookstore on February 22, 2024, which the author of the article attended.

This period in the history of object-oriented philosophy is associated with the individual efforts of Harman, to whom the image of a lone wolf fits perfectly. This image was ironically embodied in the confrontation between the Prince and the Wolf in the book of the same name (Latour et al., 2011), dedicated to the public discussion of Harman with Latour on the new metaphysics at London School of Economics (2008).[5] By the time of the discussion, Harman had not yet acquired his own "pack," but the process of forming a school had already begun.

AT THE ORIGINS OF SPECULATIVE REALISM (2007–2011)

The second period in the history of OOO is associated with the emergence of a group of Harman's followers — the classic quartet of object-oriented philosophers. But it begins with the gathering of another quartet of philosophers — the co-founders of speculative realism, in which Harman played an important role.

In April 2006, at the annual congress of the Nordic Society for Phenomenology in Iceland, Harman, in his words, had "a big fight... with the 'Husserlian mafia'" (Brassier et al., 2007: 376) about the interpretation of Heidegger's philosophy. And the search for like-minded people in the confrontation with the conservatively minded continental philosophical establishment became an urgent task for him. Shortly before this, the British philosopher Ray Brassier (b. 1965), who had invited Harman a year earlier to lecture at Middlesex University in London, drew his attention to the book *Après la finitude* (Meillassoux, 2006) by the French philosopher Quentin Meillassoux (b. 1967). The idea to hold the workshop arose almost immediately.[6] The British philosopher Ian Hamilton Grant (b. 1963) joined as the fourth participant.

The first workshop, entitled "Speculative Realism," took place on 27 April 2007 at Goldsmiths College, University of London. It was co-sponsored by the philosophical journal *Collapse*, which published a ful transcript in Volume 3 (2007) (Brassier et al., 2007). Brassier, Grant, Harman and Meillassoux (in order of presentation) were the speakers, and the Italian philosopher Alberto Toscano (b. 1977), representing Goldsmiths, served as a moderator. According to Harman, the title of the workshop was suggested

[5]Back in 1999, in one of his early works, Harman called Latour the King of Networks; in the new iteration, the King has become the Prince of Networks. In turn, the French philosopher ironically compared the critics pursuing him to wolves.

[6]Harman was in correspondence about organizing the seminar from Iceland.

by Brassier as a compromise between the participants' positions. An earlier version had been "Speculative Materialism," but this title did not suit Harman's own "ardently *anti*-materialist position" (Harman, 2018b: 13).

The second workshop of the group at the University of the West of England (April 24, 2009: Bristol, UK) already contained a reference to a certain split in its title — "Speculative Realism / Speculative Materialism." The same philosophers took part in the meeting, except for Meillassoux, whose article was presented by Toscano. The group never met again in this composition, and after some time, disagreements between its participants called into question the viability of the speculative realism movement.

Researchers and the participants of the movement themselves agree that speculative realists are substantively united by a common critical position toward the methods of the continental philosophy that preceded them, which Harman characterized as the *philosophy of human access*, Meillassoux as *correlationism*, and Brassier as the *argument of the Gem*.[7] As a result, Meillassoux's term became the most widely used among speculative realists. The movement broke up into four independent directions:

- Object-Oriented Ontology (Harman);
- Speculative Materialism (Meillassoux);
- Transcendental Materialism, or Neo-Vitalism (Grant);
- Radical Nihilism (Brassier).[8]

Eventually, speculative realism continued to exist in the form of the Harman-Grant alliance,[9] while Brassier and Meillassoux, recognizing each other as allies, separated from it.[10]

In 2011, in an interview with the Polish journal *Kronos*, Ray Brassier spoke quite harshly about speculative realism, in the creation of which he had participated:

> The "speculative realist movement" exists only in the imaginations of a group of bloggers promoting an agenda for which I have no sympathy whatsoever:

[7]See more, Gratton, 2014; Harman, 2018b.

[8]Graham Harman has called Grant's direction *Vitalist Idealism*, and Brassier's — *Prometheanism* (ibid.).

[9]By the mid-2020s, this alliance seemed to have broken down, as indicated by the implacable criticism of speculative realists by Timothy Morton, who presented them as opponents of OOO (Morton, 2024a). The original article, titled "Hideous Gnosis Unbound: The Apotheosis of Speculative Realism," was published on December 20, 2024, in Morton's blog *Bring Me My Bow of Burning Gold* and removed from there in April 2025 during a rebranding. Thus, the text remains available only in Russian translation.

[10]For various interpretations of speculative realist alliances, see Žižek, 2012: 640; Harman, 2013: 23–26.

> actor-network theory spiced with pan-psychist metaphysics and morsels of process philosophy.[11] [...] I agree with Deleuze's remark that ultimately the most basic task of philosophy is to impede stupidity, so I see little philosophical merit in a "movement" whose most signal achievement thus far is to have generated an online orgy of stupidity (cit. in: Gratton, 2014: 3).

Brassier was referring to the network of philosophical blogs through which "speculative realism gained influence and grew into a movement" (Pisarev & Morozov, 2020: 27). Of the Goldsmiths quartet, only Harman was active in them. Although the participants of the philosophical blogosphere viewed it positively as a useful communication platform that allowed them to present ideas at an early stage, provided the opportunity to receive a faster feedback from colleagues than in journal publications, and leveled academic hierarchies (Bryant et al., eds, 2011: 6–7), its downside was trolling, squabbling, and mutual insults (Gratton, 2014: 3). Brassier's skepticism had some reason.

However, it would be wrong to reduce speculative realism to blogs: gradually it began to receive more traditional forms of academic institutionalization. In 2010, the almanac *Speculations* began to be published, positioning itself as the first journal devoted to speculative realism (ed. by Paul Ennis). In January 2011, a collection entitled *The Speculative Turn: Continental Materialism and Realism* was published (ed. by Levi Bryant, Nick Srnicek, and Graham Harman). It was devoted to the discussion of the ideas of the quartet of speculative realists. In February 2011, Edinburgh University Press launched the series "Speculative Realism" (ed. by G. Harman).[12] The Polish journal *Kronos* published a thematic issue (№ 1 2012) on speculative realism with articles and Harman-Meillassoux debates. The Russian philosophical journal *Logos* published a thematic block of texts entitled "Speculative Realism" (No. 2 2013), as well as two issues of the so-called "Dark Logos" (No. 4–5 2019).

Speculative realism has been the subject of a number of monographs (Ennis, 2011; Gratton, 2014; Kozlova & Joy, 2016; Shaviro, 2014). Graham Harman attempted to survey the ideas of speculative realism three times (Harman, 2011a,b; 2013) before publishing his own comprehensive interpretation of this phenomenon in a separate monograph (Harman, 2018b). Two

[11] This caricature depiction points in the direction of Graham Harman and Ian Hamilton Grant.

[12] By early 2025, the series already included 18 monographs, the last of which was published in January 2024. See the series page: `https://edinburghuniversitypress.com/series-speculative-realism/`.

collections edited by Charlie Johns and Hilan Bensusan (Johns & Bensusan, 2024; Johns & Bensusan, 2025) aim to offer a kind of summation of more than fifteen years of the history of speculative realism.

The position of object-oriented ontology in the context of speculative realism is a kind of paradox. On the one hand, it is one of the directions of this broader movement. On the other hand, from a certain point of view, speculative realism can even be seen as a phenomenon that developed within the internal logic of OOO itself. At every fork in the road, OOO ended up prevailing in the struggle for the legacy of speculative realism. In the Harman-Grant alliance, OOO philosophers played the leading role.

Today, according to Bensusan, "What we have observed is a slow death [of speculative realism] not by crumbling, but by dissemination." (Bensusan, 2025: 290). But this cannot be said of OOO, which continues its development and expansion into other spheres of human knowledge and activity, while maintaining a certain integrity and methodological rigor.

"WE'RE MORE POPULAR THAN DELEUZE NOW..."[13]

A year after the Goldsmiths workshop, object-oriented philosophy became the subject of another high-profile public event. In February 2008, London School of Economics hosted a one-day public debate between Graham Harman and Bruno Latour on the new metaphysics, entitled "Harman's Review of Bruno Latour's Empirical Metaphysics." The discussion was focused on Harman's manuscript of *The Prince of Networks: Bruno Latour and Metaphysics* (Harman, 2009). While acknowledging that "thanks to Latour, object-oriented philosophy has become possible" (ibid.: 228), Harman also pointed out the essential differences between the two theories. This debate influenced not only Harman's work, but also Latour, who later proposed the concept of "object-oriented politics" (Latour, Porter, 2013). At that time, Harman was working to demarcate OOP from philosophical approaches that were in the same theoretical field. In addition to ANT, the closest "competitors" of OOO were Manuel DeLanda's ontology (Harman, 2008) and Quentin Meillassoux's speculative materialism (Harman, 2011a).

Shortly after the LSE debate, Harman was contacted by Levi Bryant, who became one of his early followers. Bryant graduated from Loyola University in Chicago and completed his doctorate in the philosophy of Gilles Deleuze (2004), published as a monograph, *Difference and Givenness: Deleuze's*

[13]Morton, 2024a. *The Beatles'* original line: "We're more popular than Jesus now": The Formation of the Object-Oriented Philosophers' Quartet (2008–2011).

Transcendental Empiricism and the Ontology of Immanence (2008). He even practiced for a time as a Lacanian psychoanalyst (Ennis, 2010: 64). After completing his doctorate, Bryant took up a position as professor at Collin College in Dallas, Texas, where he continues to work today.

In 2008, the philosophers began corresponding, the reason for which was the preparation of *The Speculative Turn*, which Bryant and Nick Srnicek began, inspired by Meillassoux's book and Goldsmiths. They invited Harman to become the third editor of the project. Bryant knew nothing about object-oriented philosophy and entered an email dispute with Harman, trying to clarify this theory for himself. He "came out of the tail end of that debate transformed [by Harman's ideas]" (Bryant, 2011: x). By that time, Bryant had already been actively running his blog *Larval Subjects* and subsequently made it a platform for the active promotion of ideas of object-oriented philosophy and speculative philosophy.

Another member of the OOO quartet, Ian Bogost, had been interested in object-oriented philosophy much earlier than Bryant, but had become involved in the movement somewhat later, influenced by philosophical blogs. He majored in Comparative Literature at the University of California in Los Angeles (Master's degree in 2001, PhD in 2004). In 2003, Bogost co-founded the video game company Persuasive Games LLC with Gerard LaFond and began working on critical theory of video games, combining philosophical approaches with media and technology studies, programming, game design, and related fields. Bogost turned his attention to the work of Graham Harman "perhaps half a year before the publication of *Tool-Being*" (Gratton, 2020: 111). In his first monograph, *Unit Operations: An Approach to Videogame Criticism* (2006), he applied some Harman's ideas to the analysis of video games and called OOP a "related concept" (Bogost, 2006: 5).

In 2008, Bogost took up a position as an Associate Professor in the School of Literature, Communication, and Culture at Georgia Institute of Technology (Atlanta, USA). After joining OOO, he became an organizer of the workshop on speculative realism entitled "Object-Oriented Ontology: A Symposium" in April 2010 (GIT, Atlanta). This workshop was the third in a series that began at Goldsmiths, but only Harman participated all three. Bryant and Bogost were also among the participants, as well as Steven Shaviro (b. 1954) and Eugene Tucker.[14]

[14] After the workshop. Shaviro and Tucker continued their work in speculative realism. They can be classified in the direction of Ian Hamilton Grant.

Timothy Morton was the last of the OOO quartet to join the movement. A native of London, he completed his doctoral thesis at Oxford University (UK) in 1992, after which he moved to the United States, where he taught at New York University (1993–1995) and the University of Colorado (1995–1999; 2000–2003). Morton's academic career developed in the fields of literary criticism and cultural studies. He established himself as an expert on English Romanticism, in particular the works of Percy and Mary Shelley. He studied issues of consumption, diet, the human body, and the relationship between human beings and the environment in the literature of English Romanticism.

In 2003, Morton took up a professorship at the University of California in Davis, where he shifted the focus of his research to environmentalism and ecological criticism, initially in the literature of English Romanticism, but gradually broadening the context. He came to OOO as the author of two groundbreaking monographs on eco-criticism (Morton, 2007; 2010). He learned about object-oriented approach from Levi Bryant's blog, which had reviewed the recently published *The Ecological Thought* (2010), and soon began to associate himself with this movement, actively participating in discussions and events.

In December 2010, the second OOO symposium (and the fourth since Goldsmiths) was held at the University of California in Los Angeles, under the title "Hello, Everything! Speculative Realism and Object-Oriented Ontology." It was the first time that the OOO quartet gathered in one place. Harman gave an introductory talk on the distinction between OOO and speculative realism. In September 2011, the third symposium on object-oriented ontology (OOO III) was held at the private research university New School in New York, which finally secured the institutional leadership of object-oriented philosophers among speculative realists. In addition to the quartet, Shaviro and Tucker also took part in the OOO III.

PUBLICATION OF KEY WORKS ON OBJECT-ORIENTED PHILOSOPHY (2011–2013)

The third period in the history of OOO begins with the publication of key works by members of the classical quartet. For Morton, Bryant, and Bogost, they were the first monographs in which these authors directly declared themselves to be object-oriented philosophers.

In July 2011, Graham Harman published one of his most important works, *The Quadruple Object* (Harman, 2011b). A year earlier, this text was published in French for the "MétaphysiqueS" series of Presses Universitaires

de France (ed. by Q. Meillassoux). Harman, in a condensed format, represented the basic principles of OOO, proposing a model of the quadruple object consisting of a real (RO) and a sensual object (SO), as well as real (RQ) and sensual qualities (SQ). In his model, he integrated the ideas of E. Husserl (discussing SO, SQ, RQ), M. Heidegger (SO, RO) and G. W. Leibniz (RO, RQ). Harman also placed OOO in the context of speculative realism. In particular, he pointed out the difference between his approach and panpsychism, which was developed by speculative realists of the Grant's direction (for example, S. Shaviro and others), contrasting it with the *polypsychism* of OOO.

In November 2011, Levi Bryant published his monograph *The Democracy of Objects* in the "New Metaphysics" series of Open Humanities Press (ed. by G. Harman and B. Latour). According to Bryant, "Every page of the book that follows is inspired by Harman's work, such that it is impossible to cite all the ways in which he has influenced my thinking" (Bryant, 2011: x). The central topic of the book — "the democracy of objects" — is most fully explored in the concept of *flat ontology*, a principle that denies any hierarchies of being. In 2014, Bryant moved away from the basic principles of OOO. In his second monograph, *Onto-Cartography: An Ontology of Machines and Media* ("Speculative Realism" Series), he partly returned to a Deleuzian perspective and proposed a new approach: machine-oriented ontology (MOO). The main difference between MOO and OOO is that Harman's objects do not directly interact with one another, whereas Bryant's machines, on the contrary, "can directly affect one another" (Bryant, 2014: 58).

In March 2012, Ian Bogost published his monograph *Alien Phenomenology, or What is it Like to Be a Thing?* In his approach, which he called alien phenomenology, Bogost, relying on the anti-correlationism common to OOO, extends the phenomenological concept of intersubjectivity not only to humans but to all living beings and even things. The book discusses and develops Harman's key concepts: ontography, metaphor, carpentry,[15] etc. Bogost also rethinks the concept of flat ontology and proposes the notion *tiny ontology* (Bogost, 2012).

In 2013, Timothy Morton published two monographs at once: *Hyperobjects: Philosophy and Ecology after the End of the World* and *Realist Magic: Objects, Ontology, and Causality* ("New Metaphysics" Series). A year before, Morton

[15]The subtitle of the *Guerilla Metaphisics* is "Phenomenology and the Carpentry of Things." Carpentry indicates the constructivist character of the work of the speculative philosopher.

left California and moved to Texas, where he took up the Rita Shea Guffey Chair in English at Rice University, Houston. In *Hyperobjects*, he directly calls himself an "object-oriented ontologist" (Morton, 2013a: 3). Both books are an application of the OOO to concepts previously proposed by Morton, such as hyperobjects and dark cognition (both have existed since 2010).

THE PEAK OF OBJECT-ORIENTED ONTOLOGY POPULARITY (2016–2021)

The academic activity of OOO philosophers continued to accelerate. In the period 2011–2023, Harman published 16 monographs and collections of articles (2 co-authored), Morton published eight (one co-authored), and Bogost one.[16] Harman also headed two academic series — "New Metaphysics" (with B. Latour) and "Speculative Realism." Bryant and the founding director of Punctum Book Eileen Joy, attempted to launch a journal dedicated especially to OOO.[17] Since 2019, the peer-reviewed open access journal *Open Philosophy* (one issue per year by De Gruyter Brill) has become a new platform for academic discussion of OOO ideas, where G. Harman became the editor-in-chief.[18] This journal publishes works by a new generation of philosophers who associate themselves with the ideas of OOO — Nicky Young, Arjen Kleinherenbrink, Jordi Vivaldi, and others.

In 2016, Harman returned to the United States and took up a professorship at the prestigious Southern California Institute of Architecture in Los Angeles (SCI-Arc). In August of the same year, he, along with Tim Morton, was included in the list of the top 50 best living philosophers according to the venture company *The Best Schools*. A year earlier, in 2015, Harman had been included in th top 100 most influential people in the field of art according to the influential magazine *Art Review* (Morton was included in 2016). In 2023, the publication of *The Graham Harman Reader* (ed.

[16]Of course, quantitative characteristics in philosophy cannot be considered the main argument for the success of a school. But for comparison, Harman's colleagues in speculative realism R. Brassier and I. H. Grant have not published a single new monograph or collection of articles since 2006; Q. Meillassoux published — 3.

[17]The journal *O-Zone: A Journal of Object-Oriented Studies* published only one issue in 2014, dedicated to ecology. It is difficult to judge the reasons for the unsuccessful launch of the journal. Perhaps they relate to the fact that *O-Zone* duplicated the almanac *Speculations*, published by the same publishing house *Punctum Book*.

[18]Three special issues have already been published with the general subtitle "Object-Oriented Ontology and Its Critics" (ed. by G. Harman, 2019, 2020, 2021), as well as a special issue "Towards a Dialogue between Object-Oriented Ontology and Science" (ed. by A. R. Sandru, F. G. L. Ortiz and Z. F. Mainen, 2024).

by J. Cogburn and N. Young) became a kind of recognition of Harman's merits as a systematic philosopher.

OOO's academic success quickly converted into its popularity among a wider audience of non-fiction readers. In 2015, Ian Bogost and writer and editor Christopher Schaberg launched the "Object Lessons" project about the "hidden life of ordinary things."[19] Its founding advisory board also included G. Harman, T. Morton, and others. The project was a series of non-fiction micrographs from Bloomsbury and a series of journalistic essays in the online version of one of the oldest American magazines, *The Atlantic*. Each micrograph tells the story of one object from a non-anthropocentric perspective, and the essays tell individual aspects of this story. The heroes of the books were a TV remote control, a golf ball, a drone, a refrigerator, a hotel, whale songs, potatoes, wine, the ocean, jeans, mushrooms, OK, an email, and many others. In total, the series, which continues to this day, has published more than 80 micrographs and more than 200 essays.

The British-American media giant *Penguin Random House*, which is aimed at a mass audience, included OOO in the publishing program of its division, *Pelican*, which specializes in non-fiction literature. The new "Pelican Books" series has published *Being Ecological* (Morton, 2018), and *Object-Oriented Ontology: A New "Theory of Everything"* (Harman, 2018a). Morton's book opened the series, and Harman's was the eighth one. Both authors offered the mass reader a popular and systematic presentation of their philosophical ideas, previously developed in other works. In 2021, the new *Penguin Random House* series, "Green Ideas," by *Penguin Classics*, published *All Art Is Ecological* (Morton, 2021). It became the third in the series after eco-activist Greta Thunberg and journalist Naomi Klein. The 2017, 2018 and 2021 editions marked the peak of the public presentation of the OOO ideas.

EXPANSION OF OBJECT-ORIENTED PHILOSOPHY INTO OTHER AREAS (2011–2023)

In his *Object-Oriented Ontology* (2018), Harman set the ironic, yet ambitious goal of demonstrating the potential of OOO as a new "theory of everything" that can be applied to a wide range of human thought and activity — the humanities, social and political sciences, natural science (especially in everything related to environmental issues), art, architecture,

[19] See more, https://objectsobjectsobjects.com/.

popular culture, etc. However, interest in OOO in these fields had emerged long before the book's publication.

Ecology. In May 2014, Timothy Morton gave the prestigious Wellek Lectures on dark ecology at the University of California, Irvine. He had been developing this concept in his works since 2005. In *Ecology Without Nature* (2007), Morton contrasted dark ecology with the activist strategy of deep ecology. The latter effectively deifies Nature and presents humans as parasites on the body of the planet. Morton rejects the concept of Nature and proposes instead to view the surrounding world as a symbiotic *mesh* of coexistence between humans, non-human living beings, and inanimate objects. In autumn 2014, inspired by Morton's ideas, the Dutch art collective *Sonic Acts* together with Hilde Methi, a curator from Kirkenes (Norway), and in collaboration with Norwegian and Russian partners, launched a three-year project, "Dark Ecology," in the border zone of the Norwegian and Russian Arctic. The goal of the project was a scientific, philosophical, and artistic understanding of the "intimate interconnections" of humans with other non-human beings — "iron ore, snowflakes, plankton or radiation" (Dark Ecology, 2017). Philosophers, ecologists, artists, and sound designers took part in the project.

The first expedition of "Dark Ecology" took place in October 2014 and traveled along the route Kirkenes-Nikel-Zapolyarny-Kirkenes. Morton visited the Nikel steel plant as part of a group and acted as a key speaker with a number of lectures during the project. The second expedition proceeded in November 2015, with Murmansk added to the itinerary. The key speaker at this stage was Graham Harman. The third expedition was happened in June 2016 and included a trip to Pasvik and Kirkenes, as well as the vicinity of Nikel. Each trip involved around 40 and 60 people. In 2016, Morton published the book *Dark Ecology: Towards a Logic of Future Coexistence* (Morton, 2016).

The Dark Ecology project (2014–2016) focused on the study of meshes of coexistence and their dark cognition in the context of the post-industrial landscape of the Norwegian and Russian Arctic. The idea of dark ecology continued to influence the eco-art agenda after the project ended. In spring 2018, it was included in the program of the festivals "Inversion" (Murmansk), "SALT ART" (Oslo), "Terminal B" (Kirkenes). It also became part of the international eco-art projects "Living Earth" (2018) and "Changing Weathers" (2014–2020).

Art. Aesthetics plays a crucial role in OOO as "the root of all philosophy" (Harman, 2018a: 59). It becomes one of the key types of indirect access to

real objects. This explains the keen interest of OOO philosophers, especially Graham Harman and Timothy Morton, in art as a way of understanding reality, both in its theoretical aspect and in practice. In turn, representatives of the art world reciprocated by inviting philosophers as experts and involving them in their art projects. Already in *Guerrilla Metaphysics* (2005), Harman began to develop his theory of metaphor, drawing on the early ideas of J. Ortega y Gasset. He addressed the problem of paraphrase in literature in his works on H. P. Lovecraft (Harman, 2012). The approaches outlined in several articles in 2014, which engaged the works of American art critics Clement Greenberg (1909–1994) and Michael Fried (b. 1939), were developed in the chapters on aesthetics of *Dante's Broken Hammer* (Harman, 2016), Chapter 2 of *Object-Oriented Ontology* (Harman, 2018a), and *Art and Objects* (Harman, 2020a). Harman's main thesis is: "All art is theatrical." This means that the art object (including literary metaphor), as a real object, is withdrawn from direct access and the beholder must take its place and perform it: "For this reason, artworks are all compounds that consist of an art object plus a beholder" (Cogburn & Young, 2023: 1365).

Morton expressed this theatricality in his own way. For him, all art is ecological, because "the experience of art provides a model for the kind of coexistence ecological ethics and politics wants to achieve between humans and nonhumans" (Morton, 2021: 10–11). Morton developed the aesthetics of OOO in close connection with environmental criticism in his works (Morton, 2013a,b; 2016; 2021, etc.). In parallel with this, he actively participated in various art projects. In addition to the already mentioned "Dark Ecology," the most famous of these include a number of installations by Justin Brice Guariglia — "We Are the Asteroid" and "Baked Alaska" (both 2018); "Human Kind Ness" (2019), for which Morton prepared the text. He also wrote the libretto for the opera "Time, Time, Time" (dir. by J. Walsh, 2019) and performed one of the roles in it.

OOO has become a source of ideas and inspiration for many artists. One of the first was the Polish artist Joanna Malinowska with her work "Time of Guerrilla Metaphysics" (2009). Icelandic singer Björk entered an email correspondence with Morton, which later became part of her solo exhibition at the MoMA in New York (2015) and was included in the publication *Björk. Archives* (2015). Among the artists whose works were inspired by OOO's ideas are Eduardo Navarro, Pamela Rosenkranz, Pierre Huyghe, Olafur Eliasson, among others. In 2019, American director, producer, and screenwriter Adam McKay, inspired by Morton's concept of hyperobjects, created the Hyperobject Industries studio. The director's

idea for the studio's first film, "Don't Look Up" (dir. by A. McKay, 2021), was inspired by "We Are the Asteroid."

Architecture and design. Graham Harman entered the debate on architectural theory several years before taking up his position at SCI-Arc. In June 2013, a workshop entitled "Is There an Object-Oriented Architecture?" was held at the headquarters of the Swedenborg Society in London. The discussion was organized by Joseph Bedford, director of the "Architecture Exchange," and Jessica Reynolds, co-founder of "vPRR Architects." In addition to Graham Harman, the participants included such theorists of architecture as Adam Sharr, Lorens Holm, Jonathan Hale, Peg Rawes, Patrick Lynch, and Peter Carl. In 2020, following the discussion, a collection *Is There an Object-Oriented Architecture? Engaging Graham Harman* was published (Bedford, 2020).

The discussion continued in October 2016 at the symposium "The Secret Life of Buildings," organized by the School of Architecture at University of Texas, Austin. The conference was attended by the OOO quartet, as well as Albena Yaneva, who was one of the first to apply ANT to architecture. In 2018, a collection of the same name was published (Benedikt & Beig, eds., 2018). Among other things, it contained a polemic exchange between Graham Harman and Patrik Schumacher (b. 1961), the author of the concept of *parametricism* in architecture (2008) and director of Zaha Hadid Architects.

Harman's work at SCI-Arc gave impetus to the application of OOO in architecture and design. One of the striking examples is the work of Tom Wiscombe, founding director of *Tom Wiscombe Architecture*. In addition to architectural projects, Wiscombe initiated several publications that discussed the application of OOO in architecture: *Objects of the Model World* (2021) and *Conversations on Architecture and Objects* (2021) with the participation of G. Harman, T. Morton, and others.

In 2022, Harman published a summary of the above discussions and a response to critics in *Architecture and Objects* (2022). He analyzed three major waves of influence of philosophical ideas on architectural theory over the past sixty years, associated with the names of M. Heidegger, J. Derrida, and G. Deleuze, and proposed his own approach as a fourth alternative.

In Russia, the OOO theory has been applied to design by Oleg Paschenko, a media designer, digital artist, and lecturer at HSE School of Design in Moscow (Paschenko, nodate).

Archaeology. OOO's expansion into archaeology began in 2014, when Graham Harman delivered the prestigious Haragan Lecture at Texas Tech University in Lubbock. The topic of time in OOO repeatedly came up during

the lectures. The TexTech professor of archaeology Christopher Whitmore invited Harman to discuss it separately and to examine several archaeological examples from the standpoint of OOO theory. The result of the discussion was a joint monograph by Harman and Whitmore, *Objects Untimely: Object-Oriented Philosophy and Archaeology* (Harman & Witmore, 2023). Harman took the opportunity to clarify one of OOO's major weaknesses: the problem of time. He develops an argument that time is generated by objects rather than encompassed by them, discussing the processual approaches to time of M. Heidegger, H. Bergson, A. N. Whitehead, G. Simondon and G. Deleuze, the concept of the unreality of time by J. M. E. McTaggart, and responding to criticism from P. Wolfendale, P. Gratton and A. Kleinherenbrink.

The influence of OOO on archaeology can also be found in the collection *Contemporary Philosophies for Maritime Archaeology: Flat Ontologies, Oceanic Thinking, and the Anthropocene* (2023). It contains articles by G. Harman and C. Whitmore, as well as one by one of the friendly critics of the object-oriented approach in archaeology, who helped in the work on *Objects Untimely*, the Norwegian archaeologist Bjørnar Olsen.

Religion. The expansion of object-oriented ontology into the field of religion is primarily associated with Timothy Morton, who openly declared himself a religious person. Graham Harman and Ian Bogost can be described as indifferent to religious issues, and Levi Bryant is openly hostile toward theistic religiosity.

Morton identified as Buddhist for quite a long time, he belonged to the Drukpa Kagyu school of Tibetan Buddhism and practiced Mahamudra and Dzogchen. Apparently, Morton recieved Buddhist initiations, as indicated by the presence of his sacred Tibetan name — *Gyurmë*. Having become acquainted with the object-oriented approach, he found many intersections with Buddhist philosophy. In 2010, he even set the ambitious task of combining their basic principles in a project of Object-Oriented Buddhism.

Morton turns to the ideas of the Indian philosopher Nāgārjuna (2nd–3rd centuries) in *Realistic Magic* (2013), developing the topic of indirect (vicarious) causality (along with medieval Arabic occasionalism). In the same work, Morton examines his own concept of *interobjectivity* through the Buddhist concept of *bardo* (Morton, 2013b: 177–184, 196–198). In 2015, he co-authored the book *Nothingness: Three Introductions to Buddhism* (Boon et al., 2015) with the writer and journalist Marcus Boon and the specialist in critical and cultural theory Eric Cazdyn.

In March 2023, Morton experienced a religious conversion and became a born-again Christian.[20] In his book *Hell: In Search of a Christian Ecology* (2024), he turns to an unorthodox Christian theological perspective without any strong references to Buddhism. Together with his wife, literary critic and writer, Trina Balds, he maintains a Substack blog on theological topics.

Russian theologian and philosopher Andrey Shishkov also applies the ideas of OOO to theology (Shishkov, 2021; Shishkov, 2022).

CONCLUSION

Summing up more than a quarter of a century of history of object-oriented philosophy/ontology, it can be said that despite its youth, there is no doubt that it has already established itself as an independent philosophical direction and school of thought. Its formation can be divided into three main stages, which, by analogy, may be compared to periods of human life.

In the first stage (1997–2006), Graham Harman single-handedly developed the basic principles of OOP, taking the first timid steps in a new direction with the support of his "godfather" Bruno Latour and gradually separating himself from the "parental figures" — the French phenomenologists, Alphonso Lingis and, of course, Martin Heidegger. In the short but stormy period of "puberty" (2007–2011), OOP sought to challenge the "adults" — the continental philosophical establishment, startling it with the scandalousness of philosophical blogs. During this time, the "teenage gang" of speculative realists emerged and disintegrated. It was replaced by a daring quartet of object-oriented philosophers, who also did not last long in full complement.

Then came a period of maturity (since 2011), associated with the institutionalization of OOO and its expansion into other areas of knowledge and activity — art, architecture, design, archeology, ecology, religion, etc. Object-oriented philosophers, first of all, Harman and Morton, received well-deserved recognition, although it came mainly from outside philosophical circles. A sign of maturity can also be called attention to criticism. For example, Harman published a book devoted to careful analysis and response to critics (Harman, 2020b). Largely due to Harman's charisma and organizational talent, OOO has acquired younger followers, among whom Niki Young from the University of Malta is considered the most promising.

[20]The term "born-again Christian" is commonly used in Evangelical, Baptist, and other Christian traditions where water baptism is considered insufficient for a full spiritual life. Morton calls himself this in: Morton, 2024b: XXVIII.

Continuing the analogy, one might wonder whether OOO is going through a "midlife crisis" associated with a reassessment of values. Some signs of such a crisis can be identified. Levi Bryant, having rethought his approach as a machine-oriented ontology (Bryant, 2014), has since offered no further significant development of his ideas. His blog *Larval Subjects*, which he actively maintained all these years, has remained inactive since September 15, 2022. Ian Bogost continues to publish the "Object Lessons" series, but already in his latest monograph (Bogost, 2016), which Harman characterized as "something like an OOO ethics, or at least an OOO art of living" (Harman, 2018b: 223), he appears to distance himself from the conceptual apparatus of OOO. Tim Morton, having experienced a religious conversion in 2023, reinvented himself as a Christian theologian in 2025, starting from a blank state. In April 2025, he deleted his blog *Ecology without Nature*, where he had been expounding his thoughts on various aspects of OOO (including Object-Oriented Buddhism) for a decade and a half. However, Morton still associates himself with the object-oriented approach and even promised to host the fourth OOO workshop at Rice University (the last one took place in New York in 2011) (Morton, 2024a).

Graham Harman has moved away from systematic philosophy into various specialized spheres — art, architecture, archeology, as evidenced by his publications of recent years. And although in these works, he contributes to the development of individual aspects of the general theoretical framework of OOO (e.g., to the theory of time: Harman & Witmore, 2023), a new systematic work like *Guerrilla Metaphysics* or *The Quadruple Object* is still lacking. It is hoped that such a work will be Harman's new book *Waves and Stones*, scheduled for publication in late 2025. It is quite possible that the books by Morton (2024) and Harman (2025) will mark the beginning of a new, fourth, stage in the development of object-oriented ontology.

REFERENCES

Bedford, J. 2020. *Is There an Object-Oriented Architecture? Engaging Graham Harman*. London et al.: Bloomsbury Academics.

Benedikt, M., and K. Beig, eds. 2018. *The Secret Life of Buildings*. Austin: Center for American Architecture & Design.

Bensusan, H. 2025. "Afterword. A Note on the Contemporary History of the Real: From Process Philosophy to Post-Speculative Realism." In *After Speculative Realism*, ed. by C. Johns and H. Bensusan, 289–298. London et al.: Bloomsbury Academic.

Bogost, I. 2006. *Unit Operations: An Approach to Videogame Criticism*. Cambridge (MA) and London: The MIT Press.

——— . 2012. *Alien Phenomenology, or What It's Like to Be a Thing*. Minneapolis: University of Minnesota Press.

——— . 2016. *Play Anything: The Pleasure of Limits, the Uses of Boredom, and the Secret of Games*. New York: Basic Books.

Boon, M., E. Cazdyn, and T. Morton. 2015. *Nothing: Three Inquiries in Buddhism*. Chicago and London: The University of Chicago Press.

Brassier, R., I. H. Grant, G. Harman, and Q. Meillassoux. 2007. "Speculative Realism." *Collapse* 3:307–449.

Bryant, L. 2011. *The Democracy of Objects*. Ann Arbor: Open Humanity Press.

——— . 2014. *Onto-Cartography: An Ontology of Machines and Media*. Edinburgh: Edinburgh University Press.

Bryant, L., N. Srnicek, and G. Harman, eds. 2011. *The Speculative Turn: Continental Materialism and Realism*. Melbourne: re.press.

Cogburn, J., and N. Young. 2023. *The Graham Harman Reader*. Winchester and Washington: Zero Books.

"Dark Ecology." 2017. Sonic Acts. Accessed July 16, 2025. `https://sonicacts.com/archive/dark-ecology`.

Ennis, P. J. 2010. *Post-Continental Voices: Selected Interviews*. Winchester and Washington: Zero Books.

——— . 2011. *Continental Realism*. Winchester and Washington: Zero Books.

Gratton, P. 2014. *Speculative Realism: Problems and Prospects*. London et al.: Bloomsbury.

——— . 2020. "Interviews: Graham Harman, Jane Bennett, Tim Morton, Ian Bogost, Levi Bryant and Paul Ennis." *Speculations* 1:84–134.

Grusin, R. 2015. *The Nonhuman Turn*. Minneapolis and London: University of Minnesota Press.

Harman, G. 2002. *Tool-Being: Heidegger and the Metaphysics of Objects*. Chicago and La Salle (IL): Open Court.

——— . 2005. *Guerrilla Metaphysics: Phenomenology and the Carpentry of Things*. Chicago and La Salle (IL): Open Court.

——— . 2008. "DeLanda's Ontology: Assemblage and Realism." *Continental Philosophy Review* 41:367–383.

——— . 2009. *Prince of Networks: Bruno Latour and Metaphysics*. Melbourne: re.press.

——— . 2010. *Towards Speculative Realism: Essays and Lectures*. Winchester and Washington: Zero Books.

——— . 2011a. *Quentin Meillassoux: Philosophy in the Making*. Edinburgh: Edinburgh University Press.

——— . 2011b. *The Quadruple Object*. Winchester: Zero Books.

——— . 2012. *Weird Realism: Lovecraft and Philosophy*. Winchester and Washington: Zero Books.

——— . 2013. "The Current State of Speculative Realism." *Speculations* IV:22–28.

——— . 2015. *Chetveroyakiy ob''yekt [The Quadruple Object]: metafizika veshchey posle Khaydeggera* [in Russian]. Trans. from the English by A. Morozov and O. Myshkin. Perm': Hyle Press.

——— . 2016. *Dante's Broken Hammer: The Ethics, Aesthetics and Metaphysics of Love*. Sydney (NSW): Repeater.

——— . 2018a. *Object-Oriented Ontology: A New "Theory of Everything"*. Pelican.

——— . 2018b. *Speculative Realism: An Introduction*. Cambridge: Polity.

——— . 2020a. *Art and Objects*. Cambridge: Polity.

——— . 2020b. *Skirmishes: With Friends, Enemies, and Neutrals*. Punctum Books.

Harman, G., and T. Pinho. 2020. "Interview." Philosophy Now. Accessed July 11, 2025. https://philosophynow.org/issues/139/Graham_Harman.

Harman, G., and T. Witmore. 2023. *Objects Untimely: Object-Oriented Philosophy and Archaeology*. Cambridge and Hoboken: Polity Press.

Holbraad, M., and M. A. Pedersen. 2017. *The Ontological Turn: An Anthropological Exposition*. Cambridge: Cambridge University Press.

Johns, C., and H. Bensusan. 2024. *15 Years of Speculative Realism: 2007–2022*. London and Washington: Zero Books.

——— . 2025. *After Speculative Realism*. London et al.: Bloomsbury Academic.

Kolozova, K., and E. A. Joy. 2016. *After the "Speculative Turn": Realism, Philosophy, and Feminism*. New York: Punctum Books.

Latour, B. 2013. *An Inquiry into Modes of Existence [Enquête sur les modes d'existence. Une anthropologie des Modernes]*. Trans. from the French by C. Porter. Cambridge (MA): Harvard University Press.

Latour, B., G. Harman, and P. Erdélyi. 2011. *The Prince and the Wolf: Latour and Harman at the LSE*. Winchester and Washington: Zero Books.

Mallarkey, J. 2007. *Post-Continental Philosophy: An Outline*. New York: Continuum.

Meillassoux, Q. 2006. *Après la finitude: Essai sur la nécessité de la contingence* [in French]. Paris: Seuil.

Miller, A. S. 2013. *Speculative Grace: Bruno Latour and Object-Oriented Theology*. New York: Fordham University Press.

Morton, T. 2007. *Ecology without Nature: Rethinking Environmental Aesthetics*. Cambridge (MA) and London: Harvard University Press.

——— . 2010. *The Ecological Thought*. Cambridge (MA) and London: Harvard University Press.

——— . 2013a. *Hyperobjects: Philosophy and Ecology after the End of the World*. Minneapolis and London: University of Minnesota Press.

——— . 2013b. *Realist Magic: Objects, Ontology, Causality*. Ann Arbor: Open Humanity Press.

——— . 2016. *Dark Ecology: For a Logic of Future Coexistence*. New York: Columbia University Press.

——— . 2018. *Being Ecological*. Pelican.

——— . 2021. *All Art Is Ecological*. Penguin Classics.

——— . 2024a. *Hell: In Search of Christian Ecology*. New York: Columbia University Press.

——— . 2024b. "Poganyy gnozis bez kontsa i kraya, ili Apofeoz spekulyativnogo realizma [Hideous Gnosis Unbound: The Apotheosis of Speculative Realism]" [in Russian]. Tsentr politicheskogo analiza [Center of Political Analysis]. Accessed July 11, 2025. https://centerforpoliticsanalysis.ru/position/read/id/poganyj-gnozis-bez-kontsa-i-kraja-ili-apofeoz-spekuljativnogo-realizma.

Paschenko, O. *Nigredo [Nigredo]: proyektirovaniye v chernom [Projecting in Black]* [in Russian]. In print. Egalite.

Pisarev, A., and A. Morozov. 2020. "Speculative Realism: Exit" [in Russian]. In *Spekulyativnyy realizm [Speculative Realism]* : *vvedeniye [Introduction]*, by G. Harman, trans. from the English by A. Pisarev, 7–33. Moskva [Moscow]: RIPOL klassik.

Shaviro, S. 2014. *The Universe of Things: On Speculative Realism*. Minneapolis: University of Minnesota Press.

Shishkov, A. 2021. "Kto skryvayet·sya v teni [Who is Hiding in the Shadows]: kontury temnoy ekkleziologii [The Outlines of Dark Ecclesiology]" [in Russian]. *Gosudarstvo, religiya, tserkov' v Rossii i za rubezhom [State, Religion and Church in Russia and Worldwide]* 2 (39): 61–89.

——— . 2022. "Dark Theology as an Approach to Reassembling the Church." *Religions* 13 (4): 324.

Žižek, S. 2012. *Less Than Nothing: Hegel and the Shadow of Dialectical Materialism*. London: Verso.

Shishkov A. V. [Шишков А. В.] The Short History of Development of Object-Oriented Ontology [Краткая история становления объектно-ориентированной онтологии] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 169–193.

Андрей Шишков
аспирант
Российский государственный гуманитарный университет (Москва);
orcid: 0009-0001-4270-1900

Краткая история становления объектно-ориентированной онтологии

Получено: 28.08.2025. **Рецензировано:** 10.10.2025. **Принято:** 18.10.2025.

Аннотация: Статья посвящена истории становления объектно-ориентированной онтологии (ООО) — школы постконтинентальной философии, основанной Грэмом Харманом

в конце 1990-х годов. В ней предлагается периодизация эволюции ООО, начинающаяся с ее возникновения как самостоятельного проекта Хармана (1997–2006). Второй период (2007–2011) охватывает роль ООО в движении спекулятивного реализма и формирование ее канонического квартета — Хармана, Леви Брайанта, Иена Богоста и Тимоти Мортона. Третий, текущий период (2011–по наст. время) детализирует институционализацию ООО и ее активную экспансию за пределы философии в такие области человеческой мысли и практики, как экология, искусство, архитектура, археология и религия. В заключении автор оценивает нынешнее состояние ООО, отмечая признаки внутренней диверсификации среди ее участников, и в то же время утверждая ее прочное наследие как самостоятельной и влиятельной школы мысли, которая успешно бросила вызов антропоцентрической парадигме в континентальной философии. Автор также выдвигает гипотезу о возможном начале нового, четвертого этапа в ее развитии.

Ключевые слова: Грэм Харман, партизанская метафизика, Иен Богост, Леви Брайант, объектно-ориентированная онтология, четверояким объект, спекулятивный реализм, Тимоти Мортон.

DOI: 10.17323/2587-8719-2025-4-169-193.

Alekseeva, E. A. 2025. "The Problem of Epistemic Injustice and Multi-Agent Model of Epistemic Diversity" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 194–220.

Ekaterina Alekseeva*

The Problem of Epistemic Injustice and Multi-Agent Model of Epistemic Diversity**

Submitted: Aug. 06, 2025. **Reviewed:** Sept. 01, 2025. **Accepted:** Oct. 18, 2025.

Abstract: This article examines epistemic injustice as a fundamental epistemological problem that undermines the possibility of obtaining reliable and complete knowledge. It explores various forms of epistemic injustice — including testimonial, hermeneutical, situational, inverted, and mutual — demonstrating how these phenomena manifest in medical contexts and beyond. The paper presents a multi-agent computational model implemented in NetLogo that simulates medical decision-making through Bayesian epistemology, involving four types of epistemic agents: patients, doctors, experts, and managers. The findings support the argument that epistemic diversity is not merely a social justice concern but an epistemological necessity, consistent with veritistic social epistemology (Goldman), perspectival realism (Massimi), and agential realism (Barad). The article concludes that overcoming epistemic injustice requires not only ethical correction of individual biases but also a more radical transformation of knowledge institutions to integrate diverse perspectives while maintaining a critical differentiation of epistemic competence.

Keywords: Epistemic Injustice, Epistemic Diversity, Bayesian Epistemology, Multi-Agent Modeling, Veritistic Social Epistemology, Perspectival Realism, Agential Realism, Medical Decision-Making.

DOI: 10.17323/2587-8719-2025-4-194-220.

The Concept of Epistemic Injustice and Its Extensions

The concept of epistemic injustice describes situations in which people from different social groups face recurrent obstacles to being recognized as knowers due to systemic biases, power imbalances, or structural exclusions in knowledge production and transmission. This concept was introduced in Miranda Fricker's seminal work, where she sought to describe the various ways in which epistemic exclusion occurs. According to Fricker's research, epistemic injustice does "not prompt thoughts about distributive

*Ekaterina Alekseeva, PhD in Philosophy; Associate Professor at the State Academic University for the Humanities (GAUGN) (Moscow, Russia), eaalekseeva@gaugn.ru, ORCID: 0000–0002–0006–5942.

Acknowledgements: The article was prepared at the State Academic University for the Humanities within the framework of the state assignment of the Ministry of Science and Higher Education of the Russian Federation (topic No. FZNF–2023–0004 "Digitalization and Methods of the Modern Information Society: Cognitive, Physical, Political and Legal Aspects").

unfairness in respect of epistemic goods such as information or education" (Fricker, 2007: 6); hence, it is not strictly related to the digital or cognitive divide. The term "cognitive divide" generally refers to differences in perception, understanding, or interpretation of complex information among individuals or groups. This process, unlike epistemic injustice, is not directed at the bearer of knowledge; rather, it presupposes gaps in perception on the part of the recipient. It occurs at the moment of recognizing someone as a reliable or unreliable source of knowledge, whose testimony can be taken into consideration or rejected. Epistemic injustice manifests in several ways: people are not recognized as being able to know anything; their knowledge is not recognized as reliable; their ability to know is questioned; or their knowledge is not understood (ibid.).

Fricker outlines two key variants of epistemic injustice: testimonial and hermeneutical injustice. Testimonial injustice is the situation in which a speaker's credibility is unfairly diminished due to prejudiced perceptions of their social identity, such as race or gender: "Testimonial injustice occurs when prejudice causes a hearer to give a deflated level of credibility to a speaker's word" (ibid.: 8). A certain systematic bias arises on the part of the one who assesses the reliability of the judgment: "The basic idea is that a speaker suffers a testimonial injustice just if prejudice on the hearer's part causes him to give the speaker less credibility than he would otherwise have given" (ibid.: 17). From the point of view of epistemic logic, the situation of testimonial injustice appears as an assessment of the degree of reliability of sources a and b by some observer O. Such an observer demonstrates a gradation of epistemic trust as a subjective assessment of the reliability of the source of information. This means that the same proposition, with an equal truth value, can be judged as true or false depending on who utters it, if a and b belong to different social (or other) groups.

Hermeneutical injustice is a state of affairs in which marginalized groups lack access to shared interpretative frameworks that would allow them to articulate their lived experiences: "Hermeneutical injustice occurs at a prior stage, when a gap in collective interpretive resources puts someone at an unfair disadvantage when it comes to making sense of their social experiences" (ibid.: 8). Essentially, what is at stake is that some experiences are not considered important or visible enough to be recognized as worthy of understanding and study. Hermeneutic epistemic injustice is not simply based on the fact that certain subjects are ascribed some inability to act as bearers of articulated knowledge, but on the some impossibility to play by

the hermeneutic rules of the dominant group. Here, we can recall, for example, Foucault's studies, which brought topics such as sexuality, corporeality, and marginalized mental states entered academic discourse. Hence, overcoming hermeneutic injustice serves not only the purpose of giving oppressed subjects the opportunity to declare themselves but also broadens the fields of knowledge. At the same time, contradictions obviously arise that go back to the traditional subject-object dichotomy, in which, in order to become a worthy and significant object of research, it is necessary to alienate one's own experience from oneself. As we will see later, this contradiction will also arise when attempting to solve the problem of epistemic injustice.

The phenomenological approach to the problem of testimony reveals how epistemic injustice manifests itself not only as a theoretical problem but also as a phenomenon that affects the very structure of human experience and interaction. Concerning hermeneutical injustice, it means that witnessing attempts to express "that which cannot enter into language or cannot be said because it is banished, forbidden, and removed from it, but nevertheless touches the very heart of human existence" (Heiden & Marinescu, eds., 2025: 84). This approach points to the necessity of recognizing witnessing as a process of creating a "common world" that includes experiences that initially "have no common measure," thus counteracting epistemic injustice and understanding knowledge as always a shared form of knowing.

Obviously, the concept of epistemic injustice is rooted in Foucault's concept of power-knowledge and intersects with Fuller's social epistemology, feminist epistemology, etc. However, Fricker's concept has a number of distinctive features, which we will examine further through the prism of veritistic social epistemology.

Because of the complexity of this problem, there are several accompanying concepts that have been suggested by other researchers who picked up the idea of epistemic injustice. Building on Fricker's work, K. Dotson offers the concept of *epistemic oppression* — the systemic exclusion of marginalized epistemologies (e. g., indigenous traditions) from dominant knowledge systems. Dotson also emphasizes the interrelation between several phenomena:

> Epistemic oppression, here, refers to epistemic exclusions afforded positions and communities that produce deficiencies in social knowledge. An epistemic exclusion, in this analysis, is an infringement on the epistemic agency of knowers that reduces her or his ability to participate in a given epistemic community. Epistemic agency will concern the ability to utilize persuasively shared epistemic resources within

> a given epistemic community in order to participate in knowledge production and, if required, the revision of those same resources (Dotson, 2012: 25).

Such additional concepts allow us to explicate the complex structure and intrinsic mechanisms of epistemic injustice.

Similarly, N. Berenstain's concept of *epistemic exploitation* highlights how marginalized individuals are unfairly burdened with educating others about their lived realities, framing both neglect and over-extraction as oppressive dynamics. Epistemic exploitation "occurs when privileged persons compel marginalized persons to produce an education or explanation about the nature of the oppression they face. Epistemic exploitation is a variety of epistemic oppression marked by unrecognized, uncompensated, emotionally taxing, coerced epistemic labor" (Berenstain, 2016: 570). We can argue that this concept is highly controversial in the epistemological context because it significantly increases the imbalance between knowledge and social justice in favor of the latter and also introduces into the epistemological plane only one area of knowledge, namely knowledge related to the experience of discrimination. Within such a framework, the whole scheme begins to look like this: knowledge about personal experience is always alienated; this alienated knowledge is used by privileged groups; that is, there are no and cannot be any purely epistemic goals for the integration of a certain experience into the epistemic context.

J. Medina suggested a rather radical concept of *hermeneutical death*, which represents the most extreme form of hermeneutical injustice. It occurs when "subjects are not simply mistreated as intelligible communicators, but prevented from developing and exercising a voice, that is, prevented from participating in meaning-making and meaning-sharing practices" (Medina, 2017: 41). This radical form of injustice involves the loss or severe curtailment of one's voice, the destruction of interpretative capacities, and the annihilation of one's status as a participant in meaning-making communities. Medina gives a historical example:

> A good illustration of measures that contribute to hermeneutical annihilation can be found in slave traders' practice of separating African slaves who spoke the same language to maximize communicative isolation and in US slaveholders' practice of punishing slaves caught speaking African languages. This illustrates the deliberate strategy of hermeneutic destruction: slave traders separated African slaves who spoke the same language in order to maximize their communicative isolation (ibid.: 47).

Much of the research on epistemic injustice focuses on how it manifests itself in areas where personal experience is important, such as medicine and health care. Patients frequently experience testimonial injustice when healthcare providers discount their testimonies due to stereotypes about illness affecting cognitive reliability and emotional stability:

> *Agential testimonial injustice* is generated by culturally prevalent stereotypes of ill persons, the majority of which build in negative accounts of their epistemic abilities. The ill are often stereotyped as, inter alia, cognitively impaired, overwrought, unable to "think straight," existentially unstable, anxious, morbid, and so on, due either to their condition or their psychological response to it. Such attributions are liable to prejudice how others perceive and evaluate their epistemic abilities (Carel & Kidd, 2017: 338).

Healthcare systems create knowledge asymmetries that privilege medical training over patient experience, effectively limiting epistemic authority to practitioners. On the other side, women medical students report having their knowledge and experiences discredited based on their gender (Blalock & Leal, 2023).

The literature describing epistemic injustice focuses heavily on medical applications and often fails to extend to broader natural science contexts. But it is reasonable to suggest that epistemic injustice also exists in other scientific fields. For example, scientists from non-European countries may face prejudices regarding their research results because they allegedly lack the relevant competence. At the same time, specific ways of interpreting reality inherent in non-European cultures, which could introduce a heuristic component into scientific research, may be ignored by local scientists themselves because they do not have the hermeneutic resources to integrate this component into their scientific work.

Several additional concepts of epistemic injustice reinforce each other, creating a comprehensive picture of this complex problem. Together they show that epistemic injustice is not just an ethical problem for individuals and their biases, but a deeply rooted political problem that requires not only virtues but also a radical transformation of knowledge institutions. Institutional restructuring should dismantle the epistemic hierarchies embedded in science, education, media, legal systems, healthcare, and other domains. But without understanding the diverse mechanisms through which epistemic injustice is reproduced, without studying how the systems of inclusion and exclusion work in the field of knowledge production, how in principle it is

possible to solve the problem not only at the social but also at the epistemological level, all these postulates will remain only declarative statements.

ADDITIONAL KINDS OF EPISTEMIC INJUSTICE

Let's turn to the case of epistemic injustice described in the article "Incline and Admonish: Epistemic Injustice and Counter-Expertise" (Shevchenko, 2020). In autumn 2019, several users in an online community reported complaints about the smell of sewage in Saransk, a city in Russia. A user representing one of the regional government agencies responded that the human olfactory threshold might be lower than the maximum permissible concentration of substances. According to the comments, community members interpreted this statement as implying that their experience was not credible. The key issue here was not that people doubted the accuracy of the measuring instruments, but rather that their personal, phenomenological experience was being dismissed. The government representative was effectively denying the validity of citizens' sensory experiences by prioritizing instrumental data over their lived reality.

The author compares this situation to telling a patient with chronic pain to stop seeking medical help simply because an MRI hasn't detected any pathologies. This case illustrates what the author identifies as a radical form of epistemic injustice — the "derivatization" of another person's experience, where someone's phenomenal experience is treated merely as a derivative of measured parameters rather than as a legitimate source of knowledge in itself.

What's interesting for us in this situation? In such a situation, people of different identities (for example, men and women) suffer because of common circumstances, and none of them possess critical privilege to be heard. Here, we can see that there is a distrust of the evidence and sensory experience of people from different social groups, placed, however, in shared situation of epistemic injustice. It is obvious that distrust was expressed not only toward to representatives of a more deprived stable group (for example, women) but also toward men. This suggests that in addition to stable patterns of epistemic injustice based on stable identity, there may also be a form of epistemic injustice that can be designated as *situational epistemic injustice*. It arises when an agent situationally finds themselves in a position in which he or she is not given epistemic trust, although in another situation it might well be granted. This approach to the issue of epistemic injustice resembles the concept of T. J. Spiegel (Spiegel, 2022), who believes that the class dimension should be at the center of the study of epistemic injustice instead

of race or gender identity. But still, the concept of class is more stable than that of situational epistemic injustice.

Of course, within situational epistemic injustice, we can also observe its more stable forms, which arise when representatives of deprived identities find themselves in a situation of double mistrust. In addition, there is a special kind of testimonial injustice, which is called *intragroup testimonial injustice* (Tobi, 2023). This kind of testimonial injustice occurs when members of the same marginalized community dismiss or devalue each other's credibility due to internalized biases or hierarchical divisions within subgroups. By fracturing trust from within, it perpetuates epistemic vulnerability even among those ostensibly united against systemic exclusion.

Finally, we can imagine the most controversial situation of epistemic injustice. Let us imagine that we have an actor whose judgments are distrusted not because of their marginalized, but, on the contrary, because of their privileged social position. It is assumed that such an actor is automatically biased towards, for example, representatives of a marginalized group and the knowledge, experience, etc., they exhibit. At the same time, this actor may indeed be biased, or, on the contrary, may reflect on their prejudices, for instance through an act of phenomenological reduction, and be completely unbiased. Let us call this state of affairs *inverted epistemic injustice*. This situation is connected with the general tendency to distrust expert institutions, since, on the one hand, critical approaches really reveal the bias of experts, while on the other hand, the inconsistency of expert assessments demonstrates the complexity of knowledge production practices and the problems associated with it. The most illustrative example of this situation was the Covid-19 pandemic, which clearly demonstrated many epistemic difficulties.

And here we reach a state where there is a situation of mistrust towards experts simply because they are experts. On the one hand, it is implied that an expert always occupies a privileged position based on his or her identity. But what if the expert is, for example, not a white, middle-aged heterosexual man, but a Black woman? On the other hand, expertise is questioned because there are examples of blatant political bias among experts. Mistrust generated by individual cases of bias seems to cast doubt on any institutions related to the search for knowledge, problem solving, etc. "Epistemic injustice is not one-sided, with a division between victims and the guilty, but general and mutual" (Tishchenko, 2020: 43). Thus, we can see another type of epistemic injustice, *mutual epistemic injustice*, when

the discrepancy in social and expert positions causes the actors to mistrust each other.

Thus, we have a complex and ambiguous structure of epistemic injustice, based on the diversity of epistemic actors and forms of their interrelations. Some of them will suffer more from testimonial injustice, others from hermeneutic injustice. We ultimately arrive at the conclusion that epistemic injustice can have not only a stable (based on stable identities) but also a flexible, situational character, and also lead to situations when the very possibility of expert knowledge is questioned. Therefore, we can postulate the necessity that models for overcoming epistemic injustice cannot assume simple solutions associated, for example, with a straightforward distributive allocation of epistemic trust.

THE WAYS OF SOLVING THE EPISTEMIC INJUSTICE PROBLEM

Naturally, most authors who study the problem of epistemic injustice offer various solutions to this problem. And these solutions depend heavily on which component worries the researcher more: social injustice in relation to the knowledge-producing actor or the improvement of epistemic practices.

One of the leading approaches builds on M. Fricker's already-familiar work, which proposes the development of personal intellectual virtues as a strategy for overcoming epistemic injustice (Fricker, 2007). Fricker argues that testimonial injustice can be addressed by cultivating epistemic virtues such as humility, integrity, and fairness in listeners. This requires people to actively consider the biases that distort their assessment of a speaker's testimony. For example, healthcare professionals could be trained to recognize how their stereotypes can lead to the dismissal of certain groups of patients' reports of pain, thereby addressing the trust deficit in healthcare. However, this solution is insufficient because of the deep and systemic nature of epistemic injustice.

Complementing this individual-focused approach, J. Medina has developed strategies for expanding hermeneutical resources to address hermeneutical injustice (Medina, 2017). Medina emphasizes the importance of creating inclusive interpretive frameworks that empower marginalized groups to articulate their experiences. His concept of "resistant imaginations" describes collective efforts to challenge systemic gaps in understanding, such as recognizing systemic racism. In extreme cases, Medina advocates for more radical approaches, including epistemic disobedience (using tactics like strategic lying when existing language fails to capture experiences of oppression)

and epistemic insurrection (revolting against oppressive expressive norms through counter-discourses).

Some researchers prefer to talk not about solving the problem of epistemic injustice in general but about overcoming it in certain areas, primarily in the field of health care (which is not surprising, since this is where it manifests itself most clearly). For instance, Carel and Kidd have documented how healthcare systems can address testimonial injustice through targeted training programs that help medical professionals recognize the biases that arise in assessing patient testimony (Carel & Kidd, 2017). Similarly, Sullivan has examined how legal systems can combat epistemic injustice by diversifying juries and judges, improving education on credibility biases, and ensuring that marginalized testimonies are not dismissed due to gendered stereotypes (Sullivan, 2017).

As for the question of how epistemic injustice in the process of knowledge production can be overcome, several solutions are also proposed here. First, it is about paying special attention to ensuring the visibility of non-mainstream epistemic communities. These are epistemological communities formed by marginalized groups, such as Indigenous knowledge communities, that generate counter-narratives and are often forced to resist hermeneutical marginalization. For example, African American communities creating alternative historical narratives to counter dominant Eurocentric histories exemplify this approach. A similar situation may arise in other regions where processes of the displacement of local cultures by imperial cultures have taken place.

Other studies highlight the need to work on building epistemic trust. This implies another institutional solution, according to which strengthening epistemic trust requires democratizing knowledge production and ensuring institutional recognition of marginalized epistemologies (Anderson, 2012). This might involve reforming scientific peer-review processes to include Indigenous ecological knowledge, thereby rectifying credibility imbalances, or similar measures. Concerning academic science, it is also necessary to ensure fair crediting marginalized scholars' work and to avoid their instrumentalization, such as citing Indigenous researchers as primary authors in environmental studies rather than merely as local informants (Berenstain, 2016).

From the point of view of an intersectional approach, epistemic injustice cannot be addressed without considering how race, gender, and class intersect and reinforce each other's effects (Dotson, 2014). Ignoring class or

treating identity categories as additive rather than co-constitutive perpetuates hermeneutical injustice, leaving entire dimensions of lived experience unarticulated and structurally invisible within dominant knowledge systems. Therefore, a truly intersectional epistemology demands not only pluralistic recognition, but also a radical restructuring of epistemic institutions to center the complex, overlapping realities of multiply marginalized knowers.

A less radical but also effective approach is counter-expertise, which refers to grassroots efforts by non-professional actors — such as activists, patients, or local communities — to challenge, reinterpret, or co-create scientific knowledge. Counter-experts engage with scientific facts through three primary modes: adopting existing facts to demand accountability (such as communities citing radiation studies to pressure regulators); unpacking "black boxes" by exposing the networks behind scientific claims; and initiating new knowledge creation when official science ignores emerging issues (Filatova, 2020).

These approaches collectively represent a multifaceted strategy for addressing epistemic injustice. They range from individual virtue cultivation to institutional reforms, from community-based resistance to epistemic solidarity, and from educational interventions to radical epistemic insurrections. Each pathway recognizes that overcoming epistemic injustice requires not only correcting individual prejudices but also transforming the structural conditions that produce and perpetuate credibility deficits and hermeneutical marginalization. By implementing these strategies in complementary ways, societies can move toward more just epistemic practices that value diverse ways of knowing and ensure that all voices receive the credibility they deserve.

Here, we can distinguish at least three types of strategies for overcoming epistemic injustice. Firstly, there is the most radical decolonial approach, which argues that epistemic injustice is implemented in the structure of standards of Eurocentric culture and science. Here, a rather radical option — rejecting these standards and even replacing European (originally colonial) science with indigenous and local knowledge systems — becomes conceivable. The second strategy consists of cultivating individual sensitivity to prejudices within the framework of institutionalized practices. Finally, the third type of strategy is focused on the inclusion of alternative expert opinions and counter-expertise in the system of knowledge production. But if we are not ready to reject the standards of rationality and continue to continue to affirm the effectiveness of science, then which strategy will we choose? And an even more complex and dangerous question: isn't this this type of

epistemic injustice — and can it be called injustice — a practically inevitable component of the processes of knowledge production?

EPISTEMIC INJUSTICE FROM THE POINT OF VERITISTIC SOCIAL EPISTEMOLOGY

Most concepts of epistemic injustice have focused primarily on injustice in the social sense and the epistemic aspect here has seemed more like an auxiliary way of describing another version of power imbalances and oppressive practices. Fricker emphasized this as follows: "there is nothing very distinctively epistemic about it, for it seems largely incidental that the good in question can be characterized as an epistemic good" (Fricker, 2007: 1). But let's try to shift the focus somewhat and approach the topic of epistemic injustice from a more epistemological rather than a social perspective. Moreover, it's possible that a radical rejection of attempts at epistemic interaction between bearers of privileged and marginalized knowledge, out of fear that representatives of the former will engage only and exclusively in the appropriation of someone else's experience for their own purposes, cannot ultimately benefit the deprived groups. By contrast, the adaptation of proven knowledge systems to one's own needs, or even the demand for such an adaptation, seems like a more constructive way of acting. Finally, let us assume that we are pursuing epistemic goals first and foremost, and that achieving justice also depends on achieving them.

In order to interpret the problem of social justice in a more epistemological than social key, we will consider it from the point of view of veritistic social epistemology of A. Goldman. Although the roots of the concept of epistemic injustice clearly go back to non-veritistic social epistemology, we will try to reconsider this situation. Besides, we will use the concept of perspectival realism of M. Massimi and K. Barad's agential realism.

At first glance, Goldman's and Fricker's approaches may seem to be opposites: Goldman aims to maximize truth (veritism) (Goldman, 1999), evaluating social practices by their ability to produce true beliefs, while Fricker focuses on fairness, analyzing how social prejudices and structural inequalities distort our treatment of knowledge holders. Goldman uses the definition of knowledge as justified true belief and, based on this, defines the goal of veritistic social epistemology:

> Veritistic epistemology (whether individual or social) is concerned with the production of knowledge, where knowledge is here understood in the "weak" sense of true belief. More precisely, it is concerned with both knowledge and its contraries:

> error (false belief) and ignorance (the absence of true belief). The main question for veritistic epistemology is: Which practices have a comparatively favorable impact on knowledge as contrasted with error and ignorance? (Goldman, 1999: 5).

In this sense, he softens the strictly realistic interpretation of knowledge. Fricker herself directly criticizes postmodern relativism:

> A crucial attraction of postmodernist philosophical thought was that it placed reason and knowledge firmly in the context of social power... But this turned out to be largely a vain hope, for the extremist bent in so much postmodernist writing led too often to reductionism, and the driving force behind the postmodernist spirit emerged as more a matter of disillusionment with untenable ideals of reason than any real will to bring questions of justice and injustice to bear in reason's entanglements with social power. Suspicion of the category of reason per se and the tendency to reduce it to an operation of power actually pre-empt the very questions one needs to ask about how power is affecting our functioning as rational subjects for it eradicates, or at least obscures, the distinction between what we have a reason to think and what mere relations of power are doing to our thinking (Fricker, 2007: 2).

Although Fricker is a feminist philosopher, her approach differs from the more "non-veritist" forms of feminist philosophy, which are incompatible with Goldman's approach (Pinnick, 2000). She uses feminist critiques of power and stereotypes not to reject objective truth but to defend the right of oppressed groups to engage in a rational search for truth. Although it is perhaps precisely this moderate position that has led her followers to propose more radical options for overcoming epistemic injustice.

It turns out that Fricker's and Goldman's projects are in many ways complementary. Goldman seeks to identify the social mechanisms that lead to truth. Fricker shows that one of the main obstacles to this path is bias and epistemic exclusion. Fighting epistemic injustice is thus not a distraction from the pursuit of truth but a necessary condition for achieving it. When we ignore the knowledge of a patient, a woman scientist, or an indigenous person, we do more than commit a moral error; we deprive ourselves of valuable evidence that could bring us closer to a fuller and more accurate understanding of reality. Goldman's veritist goal of maximizing truth requires the epistemic inclusivity advocated by Fricker. As Coady (Coady, 2010) notes, even if a marginalized group is unfairly denied access to certain knowledge (e. g., professional) and cannot interpret its experience in the relevant categories, it still has unique experiential knowledge. Ignoring such experience is not only a social injustice but also an epistemic error.

To justify the epistemic value of knowledge diversity, we can turn to M. Massimi's concept of perspectival realism (Massimi, 2022). Massimi offers a realistic, but not naive, view of science. She argues that scientific knowledge is produced through "modally stable phenomena"— stable events that manifest themselves in different experimental and theoretical contexts. The key idea is that different perspectives (including local, "profane" knowledge) are like different "windows onto reality," different angles of view on the same phenomenon. Local knowledge (for example, indigenous ecological knowledge) is not "pre-scientific." It represents a unique perspective that can reveal aspects of reality that are inaccessible to more "global" science. Cutting off such knowledge ("epistemic cutting") is not just an injustice; it is an epistemic misery that leads to an incomplete and distorted picture of the world. Integrating these perspectives allows science to better identify "modally stable phenomena" and build more reliable models.

Finally, the agential realism of K. Barad provides an ontological justification for the epistemic value of all forms of experience (Barad, 2005). She introduces the concept of "material-discursive practices" in which knowledge and reality co-emerge: "In fact, agential realism offers an understanding of the nature of material-discursive practices, such as those very practices through which different distinctions get drawn, including those between the 'social' and the 'scientific'" (ibid.: 201). Entities (including the knowing subject) do not exist independently but "intra-act" with each other, producing phenomena. This means that there is no neutral, objective observer. All knowledge is knowledge from a position shaped by specific material-discursive configurations. The experience of a different (not fitting into the mainstream discourse) bodily subject is not a "subjective opinion," but a legitimate form of knowledge generated by a unique configuration. When dominant systems ignore these experiences (committing, in Fricker's words, testimonial or hermeneutic injustice), they are not simply biased; they are committing an ontological error, depriving themselves of access to entire layers of reality that could be materialized through these alternative practices. Barad shows that epistemic justice is not just an ethical imperative, but an epistemological necessity. Without "ontological humility" and the inclusion of marginalized voices, our collective understanding of the world remains fundamentally incomplete and inadequate.

Thus, comparing the positions of Goldman, Fricker, Massimi, and Barad allows us to rethink the problem of epistemic injustice. It is not just a social problem that requires ethical correction of individual biases. It is a profound epistemological problem that undermines the very possibility of obtaining

reliable and complete knowledge. The struggle for epistemic justice is not a departure from the search for truth but an integral part of it. One of the most obvious and at the same time most difficult to implement approaches to solving this problem is the principle of epistemic diversity.

THE PROBLEMS OF EPISTEMIC DIVERSITY

Epistemic diversity is a broad epistemological and practical approach of including different types of epistemic agents in the field of knowledge production. It encompasses differences in how individuals or communities form beliefs, validate knowledge and approach problem-solving (Pinto & Pinto, 2023). This synthesis of perspectives illustrates that epistemic diversity is not merely about inclusion but about fundamentally reshaping how knowledge is produced, validated, and applied across different epistemic agents and communities. This approach is consistent with the principles of perspectival realism, as well as with the principle of diversity and the material foundation of epistemic practices. Moreover, if epistemic diversity is a necessary component of epistemic activity, the question arises of how to relate it to the presence of epistemic inequality, which may also turn out to be one of the components of our epistemic practices.

Epistemic diversity is possible in several interconnected forms (Leonelli, 2022). A primary form is found within disciplinary and epistemic communities, where individuals' affiliations (for instance, engineers, social scientists, or indigenous knowledge holders) shape distinct epistemic standpoints. This diversity can be analyzed across three levels: the individual (personal beliefs), the working (contexts shaped by one's occupation), and the group (shared perspectives forged through collaboration).

Closely tied to this is the critical dimension of linguistic diversity. Language profoundly influences how knowledge is articulated, structured, and validated. Monolingual environments, such as English-dominated academia, can actively suppress epistemic diversity by marginalizing non-dominant languages and the unique conceptual frameworks they carry. Conversely, ensuring linguistic equality and allowing participants to use their preferred language fosters a much richer and more authentic exchange of knowledge.

Another vital form is the diversity of cultural and traditional knowledge systems. Indigenous and local knowledges, for instance, often emphasize relational, experiential, and context-specific understanding, which can contrast sharply with more universalizing scientific paradigms. These differences in cultural practices and values fundamentally shape how knowledge is produced, shared, and deemed credible.

Finally, epistemic diversity is expressed through methodological diversity in research and inquiry. This encompasses the classic variations between quantitative and qualitative approaches, reductionist and holistic frameworks, and empirical versus interpretive paradigms. Underpinning these methodological differences are divergent validation criteria: different epistemologies prioritize different standards for what counts as valid knowledge, whether it is empirical evidence, communal consensus, or spiritual authority.

Epistemic diversity seems like an obvious way to overcome epistemic injustice, but in trying to establish it in theoretical and practical areas we inevitably encounter a number of both epistemological and organizational problems. The first of these problems is the question of the closed nature of epistemic systems, their hermeneutic opacity. The question arises of how to overcome this closed nature, when different epistemic subjects literally speak different languages. A simple solution — asserting the equivalence of epistemic systems while leaving each confined to its own local framework — does not solve the problem of increasing knowledge through diversity. After all, in the ideal case, for example, not only should the epistemic experience of Indigenous cultures be integrated into the structure of scientific research, but the Indigenous cultures themselves should adjust or recalibrate certain ideas under the influence of scientific discourse. At the same time, it is necessary to maintain the difference and disagreement between the ways of obtaining knowledge, otherwise the necessary diversity will not be observed.

Another point is associated with the risk that the epistemic exploitation and alienation of knowledge from the personal experience of its bearer will not disappear. For example, residents of countries with a sufficiently high level of digitalization can constantly voluntarily agree to have their personal digital monitoring data on their health (pulse, number of steps taken, etc.) transferred to specialized campaigns. Moreover, they can take part in specialized surveys on their personal experience, that is, they literally provide evidence of the states they experience, etc. Yet how this data will be used and what kinds of research results will be produced on its basis, remains largely opaque to them (Zuboff, 2019).

One more problem is related to the question of whether the principle of epistemic diversity will not, on the contrary, lead to a situation of inverted epistemic injustice. Distrust of scientific expertise, or the inclusion of epistemic actors in any contexts solely on the basis of the principle of epistemic diversity, may not be justified in all contexts. As some studies show, in some tasks a homogeneous group of experts copes better than a group

observing the principle of diversity, whereas in other context the opposite may well be true.

Finally, is it possible in principle to reach some agreement on the truth of a certain judgment in groups where the positions initially do not converge and all of them have absolutely equal epistemic weight? Should epistemic trust be distributed equally, or is there a need for differentiation based on areas of competence and the degree of trust in epistemic agents. To address this question, let us turn to the possibilities offered by computer modeling of a specific epistemic case.

COMPUTER MODELING IN SOCIAL EPISTEMOLOGY

To build the model of epistemic diversity we will use a multi-agent computer simulation. Multi-agent modeling (or multi-agent system modeling) is a computational modeling approach used to simulate the actions and interactions of multiple autonomous agents (which can represent individuals, groups, organizations, or even abstract entities) within an environment. The primary goal is to understand how the behavior and decisions of these individual agents, often following relatively simple rules, give rise to complex, emergent system-level phenomena, patterns, or outcomes. Multi-agent modeling is becoming a fairly common modeling method in contemporary computer epistemology. NetLogo is one of the most useful programmable environments for modeling complex systems consisting of many actors and changing over time.

Multi-agent modeling is a relevant method of epistemological research, since

> modeling human-machine interactions, as well as the implementation of multi-agent systems, can be improved using philosophical approaches, while philosophical hypotheses about cooperation and collective activity can be tested by implementing them in artificial systems (Misselhorn, 2015: 3).

Multi-agent modeling, as the most noticeable trend in the field of computer epistemology, is most applicable within the framework of the social-epistemological approach. This is due to the fact that social epistemology studies processes that lend themselves well to interpretation by means of multi-agent systems: group beliefs, the dynamics of discussion and agreement, the formation of shared conclusion, or the distribution of confirmation across many actors, etc.

Let us consider how multi-agent modeling is used to study problems related to the topic of epistemic diversity. The first research "Yes, No, Maybe

So: A Veritistic Approach to Echo Chambers Using a Trichotomous Belief Model" (Baumgaertner, 2014) is devoted to the problem of echo chambers effect. This effect is based on the fact that the opinions of actors confirm and reinforce each other in a closed community; such a situation creates a unified internal position on certain issues and does not allow alternative information to enter into the system. As a result, actors become uniform in their abilities to perceive and produce information, which, accordingly, leads to a decrease in intra-group diversity, and this is considered as a negative phenomenon, because it suppresses the diversity of views needed to make good decisions. The study proposes to evaluate an alternative strategy of impartiality, that is, seeking interaction with different people, where agents allow their opinions to be shaped by others.

The study uses agent-based modeling with NetLogo: agents randomly move through space and interact when they encounter each other, updating their beliefs and their embeddedness according to specified rules. The key findings are that the practice of impartiality (random encounters between agents in the model) alone is insufficient to reliably prevent or destroy echo chambers. The ability to mitigate this effect depends on the embeddedness of beliefs, including false ones. Therefore, the lower the embeddedness of agents' beliefs, the greater the likelihood that the echo chamber effect will be destroyed. On the other hand, low embeddedness of the beliefs of some agents (capable of impartiality) indicates that such agents lack a stable position and may subsequently succumb to the influence of agents with more embedded beliefs. Taken together, this demonstrates that the problem of establishing diversity in the practices of acquiring and producing knowledge and information is even more complex than it might initially appear:

> ...sometimes the echo chamber effect may even be desirable under certain conditions. For example, if some agents are "designated" as possessing the truth, then all other agents may have to join them... However, if there are enough "dissenters" around (ordinary agents with sufficiently ingrained attitudes opposite to the designated agents), then they can counteract the effects of the designated agents by meeting each other between meetings with the designated agents. If agents could distinguish between each other and thus control whom they meet, this would provide a way to maintain dissent. But this is precisely what the practice of impartiality is intended to avoid (ibid.: 2562).

L. Hong and S. Page mathematically prove that, under certain conditions, a collective of randomly selected but functionally diverse agents outperforms a group of the best individual performers in their paper "Groups of

Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers" (Hong & Page, 2004). The authors model solvers as agents with unique "perspectives" (ways of conceptualizing the problem) and "heuristics" (solution-finding algorithms). The key finding — that "diversity trumps ability" — is based on the fact that in a large population, the "best" agents, selected based on individual performance, become too similar in their approaches and tend to get stuck in the same local maxima. In contrast, a random group, thanks to a greater diversity of strategies, covers a wider search space and is more likely to find the global optimum. As the authors note, "Their relatively greater ability is more than offset by their lack of problem-solving diversity" (ibid.). Computer experiments confirm this: for example, with parameters $l = 12$ and $k = 3$, a group of 10 random agents achieved a collective score of 94.53% versus 92.56% for the top 10, while the corresponding average diversity was 90.99% versus 70.98%.

This result has profound practical and theoretical implications. The authors emphasize that an individual's value in a team is determined not so much by their absolute ability to solve a problem alone, but by the uniqueness of their approach relative to other team members. They write:

> Thus, even if we were to accept the claim that IQ tests, Scholastic Aptitude Test scores, and college grades predict individual problem-solving ability, they may not be as important in determining a person's potential contribution as a problem solver as they would be measures of how differently that person thinks (ibid.).

The article offers a powerful argument for encouraging functional diversity in organizations, especially in innovative industries, where success depends on the continuous search for new solutions. The authors clearly distinguish between functional diversity (differences in thinking) and identity diversity (race, gender, age), warning that the former does not always correlate with the latter, and that the latter can create communication barriers. In conclusion, they call for the creation of organizational structures that "take advantage of the power of functional diversity" and even suggest encouraging "even greater functional diversity, given its advantages" (ibid.).

In the other research, "Diversity, Ability, and Expertise in Epistemic Communities" (Diversity, Ability, and Expertise..., 2019), the authors show that the success of the results is compared across different "epistemic environments" (landscapes). In particular, on smoother (structured) landscapes, where success on one task predicts success on another, expert groups (highly skilled agents) begin to outperform diverse groups. The authors argue that agents who perform "best" on a single, randomly generated problem should

not be considered true experts. Genuine expertise implies the transferability of skills as the ability to perform well not just on one isolated task, but across a range of related problems within a domain. In the original Hong and Page model, problems are represented as "random landscapes," where the value (or correctness) of a solution at any one point has no correlation with the values at neighboring points. In such an environment, an agent's success on one problem is purely coincidental and offers no predictive power for success on another. The "best" agents in these simulations are therefore not experts; they are merely "lucky" individuals whose specific problem-solving strategies happened to align perfectly with the arbitrary structure of that one particular landscape. To create a more realistic model, the authors introduce a parameter called "smoothness." A "smooth" landscape is one where the value of solutions at nearby points is correlated; in other words, good solutions tend to be clustered together. This mirrors real-world problems, where effective strategies (like "hill-climbing") can be incrementally refined and applied to similar challenges.

We are interested in two aspects of these articles. First, the idea of using an epistemic landscape allows us to model not just some agreement among agents regarding the truth of judgments, but also the search for the best solution to a problem. Second, the demonstration that achieving epistemic diversity is a rather complex solution and requires not just the equality of epistemic agents but also close attention to the situational context.

The potential of computer modeling as a basis for philosophical, including epistemological, research even leads to the conclusion that computer modeling should become the main philosophical method (Mayo-Wilson & Zollman, 2021). It is assumed that computer simulations, primarily multi-agent models that study interactions between actors , should be used as a more accurate alternative to thought experiments. For example, one can imagine how the thought experiment "Trolley Problem" is modeled at the level of a multi-agent model, where agents "make a decision." The authors argue that

> simulations should not displace other philosophical methods. Rather, simulations should be a tool in the philosopher's toolbox that can be used along with thought experiments, careful analysis of arguments, symbolic logic, probabilistic analysis, empirical research, and many other methods. But simulations are particularly useful in several philosophical fields, including social epistemology, social and political philosophy, and philosophy of science (ibid.: 32).

Thus, computer simulations may eventually become a key component of social epistemological research. Another question is whether the accuracy

of simulations is a complete substitute for the imprecise assumptions of thought experiments, or whether both approaches should continue to exist in parallel in epistemological research.

THE AGENT-BASED MODEL OF SPECIFIC EPISTEMIC SITUATION

One of the key questions concerning epistemic diversity: can representatives of different epistemic groups come to some common decision that can be recognized as the best possible decision of some problem? This complex issue forces us to raise a critical question: when we recognize something as rather reliable knowledge, but it is in contradiction with everyday group (or personal) experience and tends to ignore this experience should we reject such knowledge? Or should we think of this knowledge as potentially repressive and consider the epistemic equality of different ways of knowing and different types of "knowledge"? One of the most illustrative examples of such situation is discussion about evidence-based medicine and making decisions about the relevant method of treatment, the effectiveness of a drug, etc., based on some consensus of interested actors.

Evidence-based medicine is a general term for an approach to medicine in which medical practice is based on the results of studies of the efficacy and safety of drugs, treatments, etc., carried out according to the principles of scientific rationality and using relevant scientific methodology. There are a number of research methods that comply with the principles of evidence-based medicine (for example, a placebo-controlled study), but there is no need to dwell on them in detail at this stage. This kind of medicine is rooted in the principles of colonial European science, rejects the everyday peoples experience, and is supported by many (but not all) governments, international organizations, and farm companies, etc.

At the same time, there is widespread rejection of evidence-based medicine among both among medical specialists — who on the one hand must adhere to clinical recommendations, and, on the other hand, rely on their own professional experience — and among patients. At the same time, clinical recommendations accepted by specialists in the field of medical management are themselves far from always grounded in the principles of evidence-based medicine; but there we are already talking about the influence of external factors, such as the distribution of resources. The rejection of the principles of evidence-based medicine by patients is often based on ignoring their personal experience and an insufficient level of epistemic trust in them. Here we are faced with an example of situational epistemic injustice.

Let us try to model a situation in which we have four types of epistemic agents interacting, who are in a situation of making a common decision, which is the best possible solution to a problem (for example, the best therapy option for treating a certain disease). Each type of agent has specific competence in their own domain.

(1) Patients — have knowledge of their own bodily experience but do not have specialized medical expertise.
(2) Doctors — have specialized medical knowledge and clinical experience.
(3) Experts — have knowledge of the results of medical research conducted using evidence-based methodology.
(4) Managers — have management experience and organizational interests, and are guided by expert opinion as well as resource constraints.

It should be noted that at least agents of types 1 and 2 can have characteristics of marginalized groups in a situation of epistemic injustice, therefore bias must be included as one of the model's parameters. It is assumed that in a situation where the degree of competence in several significant parameters of each type of agent is taken into account, continuous interaction and mutual adjustment of their local knowledge becomes possible.

We propose an agent-based model implemented in NetLogo 6.4 that simulates generalized medical decision-making as an epistemic situation through the lens of Bayesian epistemology. The model integrates dynamic belief updating with considerations of epistemic injustice, aligning with the formal framework for integrating diverse information sources and updating degrees of belief that Bayesian models provide. This type of model can also be applied to other epistemological situations with appropriate correction.

Interaction between agents occurs within an epistemic landscape, which represents a space of possible solutions, each characterized by a value $V(x)$, which determines its effectiveness or quality. The goal of the agents is to find the global maximum of this function, i.e., the best solution x^*, for which $V(x) = maxV(x)^*$. In computational experiments, the landscape is modeled as a random function, where $V(x)$ for each $x \in X$ is generated independently from a uniform distribution on the interval $[0,\ 100]$.

Each agent has a degree of trust, both in itself and in its area of expertise. This level of trust influences the weight with which information received from one agent is considered by another. This allows us to model the differentiated perception of information depending on role, reputation, and area of expertise. During interaction, agents exchange decisions and update their beliefs based on a Bayesian approach, in which new data (in this case, decisions and their values) update the probabilities that a given

decision is the best one. This simulates rational belief updating, in which the credibility and authority of the source play a key role. The model takes into account the competence of agents, which influences the accuracy of information exchanged during interaction, and allows agents to make more informed decisions based on trust and information quality.

The agent interaction mechanism is implemented as a probabilistic process in which an agent selects a random partner, compares the quality of the partner's proposed solution with its own, and updates its belief if its trust in the partner and their area of expertise is sufficiently high. This approach reflects real-world knowledge exchange processes, where people and professionals argue, discuss, and adjust their beliefs based on trust, authority, and credibility. The model includes the ability to configure the number of agents of each type, the level of trust in them and their areas of expertise, and the number of simulation steps. This allows us to explore how different trust configurations influence belief dynamics and the achievement of consensus among agents on the best solution. The goal of the model is to demonstrate the conditions under which the maximum number of agents converges on the global maximum of the epistemic landscape — that is, the best solution — and achieves consensus.

During the experiment, we will consider three situations of modeling the level of trust in agents interactions. It is possible to flexibly configure the values of parameters such as trust in agents in NetLogo. We will use 100 agents of each type in the model, which interact with each other and with agents of other types over 300 iterations.

The first situation is that of limited epistemic diversity. In the model, this denotes a fairly high degree of trust in each agent in his area of competence, and at the same time preserving clear differences in the level of trust in each of the areas.

	personal experience (0.5)	clinical experience (0.7)	evidence-based medicine (0.9)	distribution of health care resources (0.3)
a_1 (Patient)	0.9	0.1	0.1	0.1
a_2 (Doctor)	0.1	0.9	0.5	0.3
a_3 (Expert)	0.1	0.5	0.9	0.2
a_4 (Manager)	0.1	0.3	0.2	0.9

Table 1. The situation of the limited epistemic diversity

The study found that, for given levels of trust in domains of knowledge and agents (0.9 in evidence-based medicine and 0.9 in experts in their field) and a population of 100 agents of each type (patients, physicians, experts, managers), approximately 75–85% of agents (primarily experts and physicians) reached agreement on the best solution to the epistemic landscape within 300 iterations. Patients and managers, possessing high trust in their domains but having lower relevance to identifying the optimal solution, had a limited impact on convergence. This leads to the conclusion that maximum consensus efficiency is achieved under the conditions of dominant expert influence in evidence-based medicine, whereas trust in personal experience and resource management does not contribute to convergence to the best solution.

The second situation is that of epistemic injustice. In the model, this denotes a fairly high degree of trust in each agent within their area of competence, while at the same time introducing systematic differences in the level of trust assigned to each area of expertise.

	personal experience (0.1)	clinical experience (0.7)	evidence-based medicine (0.9)	distribution of health care resources (0.3)
a_1 (Patient)	0.1	0.1	0.1	0.1
a_2 (Doctor)	0.1	0.1	0.1	0.1
a_3 (Expert)	0.9	0.9	0.9	0.9
a_4 (Manager)	0.5	0.5	0.5	0.5

Table 2. The situation of epistemic injustice

In the second model, constructed under conditions of epistemic injustice — when patients are distrusted even in their area of expertise (personal experience) — the system exhibits a fundamental cognitive dysfunction: although physicians, experts, and managers actively interact and can formally converge on a point that maximizes a given alignment score, this "highest point" proves illusory, as it is based on incomplete data and an artificially low weight of personal experience (0.1). Without access to genuine patient knowledge, professionals optimize decisions within an information bubble, which leads either to a local rather than a global maximum, or to divergence within the group of "authoritative" agents themselves — especially if their initial preferences differ and patient feedback is absent. As a result, not all 300 professionals reach even formal consensus (approximately 65–70% of

agents), and the truly best solution — one that reflects reality — remains unattainable for the entire system. Thus, epistemic injustice not only excludes patients from the cognitive process, but also undermines the rationality of the expert subsystem itself, making complete and meaningful convergence to the true optimum impossible.

The third case describes a situation of epistemic equivalence across all domains and all agents without any restrictions. This is a situation in which expertise loses its significance, as every judgment becomes equivalent to an expert judgment.

	personal experience (0.9)	clinical experience (0.9)	evidence-based medicine (0.9)	distribution of health care resources (0.9)
a_1 (Patient)	0.9	0.9	0.9	0.9
a_2 (Doctor)	0.9	0.9	0.9	0.9
a_3 (Expert)	0.9	0.9	0.9	0.9
a_4 (Manager)	0.9	0.9	0.9	0.9

Table 3. Epistemic equivalence

With uniform trust (0.9 across all agents and domains), when expertise is not taken into account, the model loses the ability to discern the reliability of information, leading to a blending of beliefs, slower convergence, and a reduction in the number of agents achieving the best solution. This demonstrates that differentiated trust is a key factor for effective agent interaction and rational agreement.

As we can see, a situation in which limited epistemic diversity exists allows for a better solution to the epistemic problem in a greater number of cases. A situation of epistemic injustice, however, reduces this figure, but not significantly. This may indirectly indicate that some degree of inequality may be present in epistemic practices without significantly reducing their success. At the same time, the absolute absence of inequality makes it difficult to find better solutions in principle. However, it should be understood that this model is highly simplified. Ideally, one should model the interaction between unique agents with a larger number of variable parameters. In that case, the results of such modeling could help reveal the more complex structure of epistemic diversity and illuminate pathways for overcoming epistemic injustice.

REFERENCES

Anderson, E. 2012. "Epistemic Justice as a Virtue of Social Institutions." *Social Epistemology* 26 (2): 163–173.

Barad, K. 2005. "Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter." In *Materialität denken : Studien zur technologischen Verkörperung — Hybride Artefakte, posthumane Körper*, ed. by C. Bath, Y. Bauer, D. W. von Bock, et al., 187–216. Bielefeld: Transcript Verlag.

Baumgaertner, B. 2014. "Yes, No, Maybe So: A Veritistic Approach to Echo Chambers Using a Trichotomous Belief Model." *Synthese* 191:2549–2569.

Berenstain, N. 2016. "Epistemic Exploitation." *Ergo: An Open Access Journal of Philosophy* 3:569–590.

Blalock, A. E., and D. R. Leal. 2023. "Redressing Injustices: How Women Students Enact Agency in Undergraduate Medical Education." *Advances in Health Sciences Education: Theory and Practice* 28 (3): 741–758.

Carel, H., and I. J. Kidd. 2017. "Epistemic Injustice in Medicine and Healthcare." In *The Routledge Handbook of Epistemic Injustice*, ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr., 32–35. London and New York: Routledge.

Coady, D. 2010. "Two Concepts of Epistemic Injustice." *Episteme* 7:101–113.

Dotson, K. 2012. "A Cautionary Tale: On Limiting Epistemic Oppression." *Frontiers: A Journal of Women Studies* 33:24–47.

——— . 2014. "Conceptualizing Epistemic Oppression." *Social Epistemology* 28 (2): 115–138.

Filatova, A. A. 2020. "Counter-Expertise: Opening and Closing the Black Boxes." *Epistemology & Philosophy of Science* 57:48–57.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Goldman, A. I. 1999. *Knowledge in a Social World*. New York: Oxford University Press.

Grim, P., D. J. Singer, A. Bramson, et al. 2019. "Diversity, Ability, and Expertise in Epistemic Communities." *Philosophy of Science* 86 (1): 98–123.

Heiden, G. J. van der, and P. Marinescu, eds. 2025. *The Phenomenology of Testimony: From Inner Truth to Shared World*. Boston: Brill.

Hong, L., and S. E. Page. 2004. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers." *Proc. Natl. Acad. Sci. U. S. A.* 101 (46): 16385–16389.

Kidd, I. J., J. Medina, and G. Pohlhaus Jr., eds. 2017. *The Routledge Handbook of Epistemic Injustice*. London and New York: Routledge.

Leonelli, S. 2022. "Open Science and Epistemic Pluralism: Friends or Foes?" *Philosophy of Science* 89 (5): 991–1001.

Massimi, M. 2022. *Perspectival Realism*. New York: Oxford University Press.

Mayo-Wilson, C., and K. J. S. Zollman. 2021. "The Computational Philosophy: Simulation as a Core Philosophical Method." *Synthese* 199:3647–3673.

Medina, J. 2017. "Varieties of Hermeneutical Injustice." In *The Routledge Handbook of Epistemic Injustice*, ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr., 41–52. London and New York: Routledge.

Misselhorn, C. 2015. "Collective Agency and Cooperation in Natural and Artificial Systems." In *Collective Agency and Cooperation in Natural and Artificial Systems : Explanation, Implementation and Simulation*, ed. by C. Misselhorn, 3–24. Cham: Springer International Publishing.

Pinnick, C. L. 2000. "Veritistic Epistemology and Feminist Epistemology: A-rational Epistemics?" *Social Epistemology* 14 (4): 281–291.

Pinto, M. F., and D. F. Pinto. 2023. "Epistemic Diversity and Industrial Selection Bias." *Synthese* 201 (5): 1–18.

Shevchenko, S. Y. 2020. "Incline and Admonish: Epistemic Injustice and Counter-Expertise." *Epistemology & Philosophy of Science* 57:20–32.

Spiegel, T. J. 2022. "The Epistemic Injustice of Epistemic Injustice." *Social Epistemology Review and Reply Collective* 11 (9): 75–90.

Sullivan, S. 2017. "On the Harms of Epistemic Injustice." In *The Routledge Handbook of Epistemic Injustice*, ed. by I. J. Kidd, J. Medina, and G. Pohlhaus Jr., 205–212. London and New York: Routledge.

Tishchenko, P. D. 2020. "Epistemic Injustice as Systemic Communicative Dysfunction." *Epistemology & Philosophy of Science* 57:42–47.

Tobi, A. 2023. "Intra-Group Epistemic Injustice." *Social Epistemology* 37 (6): 798–809.

Zuboff, S. 2019. *The Age of Surveillance Capitalism*. New York: Public Affairs.

Alekseeva E. A. [Алексеева Е. А.] The Problem of Epistemic Injustice and Multi-Agent Model of Epistemic Diversity [Проблема эпистемической несправедливости и многоагентная модель эпистемического разнообразия] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 194–220.

Екатерина Алексеева
к. филос. н., доцент
Государственный академический университет гуманитарных наук (Москва);
orcid: 0000–0002–0006–5942

Проблема эпистемической несправедливости и многоагентная модель эпистемического разнообразия

Получено: 06.08.2025. **Рецензировано:** 01.09.2025. **Принято:** 18.10.2025.

Аннотация: В данной статье эпистемическая несправедливость рассматривается как фундаментальная эпистемологическая проблема, подрывающая возможность получения достоверного и полного знания. Автор рассматривает различные формы эпистемической несправедливости, включая свидетельскую, герменевтическую, ситуативную, инверти-

рованную и взаимную, демонстрируя, как эти явления проявляются в медицинском контексте и за его пределами. В статье представлена многоагентная вычислительная модель, реализованная в NetLogo, которая имитирует принятие медицинских решений посредством байесовской эпистемологии с участием четырех типов эпистемических агентов: пациентов, врачей, экспертов и менеджеров. Результаты подтверждают аргумент о том, что эпистемическое разнообразие — это не просто проблема социальной справедливости, а эпистемологическая необходимость, согласующаяся с веристской социальной эпистемологией (Голдман), перспективным реализмом (Массими) и агентным реализмом (Барад). В статье делается вывод о том, что преодоление эпистемической несправедливости требует не только этической коррекции индивидуальных предубеждений, но и более радикальной трансформации институтов знания для интеграции различных точек зрения при сохранении критической дифференциации эпистемической компетентности.

Ключевые слова: эпистемическая несправедливость, эпистемическое разнообразие, байесовская эпистемология, многоагентное моделирование, веристическая социальная эпистемология, перспективный реализм, агентный реализм, принятие медицинских решений.

DOI: 10.17323/2587-8719-2025-4-194-220.

Kravtsov, A. D. 2025. "Morality without a Subject: Confucian-Buddhist Foundations of Ethics in the Japanese Translation of Dostoevsky's 'Crime and Punishment'" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 221–241.

Andrei Kravtsov*

Morality without a Subject**

Confucian-Buddhist Foundations of Ethics in the Japanese Translation of Dostoevsky's "Crime and Punishment"

Submitted: Sept. 06, 2025. **Reviewed:** Oct. 26, 2025. **Accepted:** Nov. 01, 2025.

Abstract: This article investigates the phenomenon of literary cultural transfer as a complex process of semiotic adaptation, where linguistic structures intersect with profound ontological paradigms. Focusing on the Meiji-era (1868–1912) Japanese translation of Fyodor Dostoevsky's *Crime and Punishment,* the study examines the mechanisms through which the Christian-existential themes of the original text are transformed under the influence of Buddhist-Confucian syncretism. The analysis centres on semantic clusters — "suffering," "conscience," and "fate" — and their ontological recoding: from Christian providentialism to Buddhist teachings on emptiness, and from existential reflection to Confucian ethics of duty. Methodologically, the framework combines corpus analysis with comparative philosophy, introducing the concepts of semantic density and cultural index as quantitative markers of axiological priorities. The author demonstrates how Russian existentialism, when confronted with the Zen concept of nothingness, generates hybrid forms: Raskolnikov's "despair" is reinterpreted through resignation, Christian "conscience" morphs into the Confucian innate virtue of *ryōshin*, and the novel's linear temporality dissolves into the cyclical model of impermanence. The philosophical significance of this research lies in its revelation of translation as a creative act that constructs new philosophical realities, where dialogue occurs not through superficial borrowings but via profound semantic metamorphosis. The translational practices of the Meiji era emerge as a space for birthing hybrid ontologies, reflecting Japan's modernization through the synthesis of traditional values and Western influences.

Keywords: Cultural Transfer, Semantic Clusters, Comparative Philosophy, Corpus Linguistics, Hybrid Ontologies.

DOI: 10.17323/2587-8719-2025-4-221-241.

The Japanese translation of F. M. Dostoevsky's novel *Crime and Punishment* during the Meiji Restoration emerges as a striking and underexplored example of such interaction, demonstrating how the Christian-existential concerns of the original are transformed under the influence of a Buddhist-Confucian worldview. This study is grounded in an analysis of semantic clusters and their transformations within translational discourse.

*Andrei Kravtsov, PhD Student at the RAS Institute of Philosophy (Moscow, Russia), drewkrav@gmail.com, ORCID: 0009–0006–1065–6136.

The Meiji period (1868–1912) was marked by a paradoxical conjunction of enforced Westernization and the preservation of traditional value foundations; this historical stage was characterized by engagement with Western thought through the prism of the Confucian ethos. This interaction became particularly pronounced in the spheres of religion and education, especially following the lifting of the ban on Christianity in 1873. By the 1880s, Japanese society was experiencing a crisis stemming from the absence of a unified "moral standard," which provoked the intensive "Debates on Moral Education" (1887–1890) (Gavin, 2004: 323). The debate was initiated by the influential scholar Kato Hiroyuki, who proposed the introduction of religious education in schools as the basis for morality, arguing that only faith in the supernatural could effectively influence the emotions of the masses (Lin & Lu, 2019: 39). This approach was opposed both by advocates of a secular, rational ethics modeled on Western exemplars and by proponents of a national morality centered on the cult of the emperor, notably advocated by Nishimura Shigeki.

The pluralistic debate was interrupted by an authoritarian state decision: on October 30, 1890, the Imperial Rescript on Education was issued, establishing a single moral standard mandatory for all, based on absolute loyalty to the emperor and Confucian virtues. This document brought intellectual inquiries to an end and, for many decades, became the ideological foundation of Japanese nationalism and militarism, defining the content of moral education (*shūshin*) until 1945 (ibid.).

From that time on, translation activity became an instrument of cultural mediation: on the one hand, it facilitated the assimilation of Western philosophical concepts, and on the other, it served as a mechanism for protecting national identity according to the principle of *wakon yōsai* (Japanese spirit — Western technology) (Wakabayashi, 2012: 180).

Reception as a philosophical-cultural phenomenon represents not a passive transmission of ideas but a complex and active process of adaptation and transformation, during which the source material is inevitably reshaped under the influence of the receiving environment. This process is often mediated by various mechanisms — whether a mediating language or the interpretive work of another thinker, which function as filters shaping the final reception. Examples from the history of Dostoevsky's reception in Europe vividly illustrate this dynamic, showing how his literary and philosophical legacy was perceived through the prism of prevailing aesthetic and intellectual norms.

The phenomenon of reconstructive reception as a process of actively transforming of foreign cultural material was not unique to the East and,

in particular, to Japan. One prominent example is the reception of Russian literature in Italy at the end of the nineteenth century, where the phenomenon of indirect translation played a key role. The first Italian translation of *Crime and Punishment* (1889) was rendered not from the Russian original but from the 1884 French translation. Despite the stigma attached to such practice, it proved decisive for the introduction of works from distant cultures, since France at that time served as the principal conduit for Russian literature in Europe (Uccello, 2024: 75). In this context, the mediating language acted as an agent of simplification and adaptation, dictated by the aesthetic preferences of the target audience. To Italian and French readers, Dostoevsky's "nervous and fragmented style" seemed devoid of the necessary "measure." Consequently, translators deemed it necessary to impose a certain "sense of proportion" on the text, which manifested in the deliberate removal of the original's linguistic tension. This process led to significant semantic and stylistic losses which, paradoxically, contributed to a more favorable reception of the work. Thus, the Italian title *Delitto e castigo* is a calque of the French *Le Crime et le Châtiment*, thereby losing the legal nuance of the Russian word *nakazanie* (ibid.: 82). Moreover, the language of the characters was standardized: social dialects and speech features that Dostoevsky employed to create psychological portraits vanished in favor of a uniformly high literary register. Thus, the reception was conditioned and shaped by the prism of a mediating agent that, through its averaging and stylistic smoothing, ensured the text's access to a new cultural environment — albeit at the cost of distortions.

Another equally significant aspect of reception emerges within the domain of European philosophical-religious thought, where Dostoevsky's ideas exerted an "almost revolutionary" influence on twentieth-century Protestant theology. His impact, particularly on Karl Barth, was likewise mediated, but not through language; instead, it occurred through the intellectual work of another scholar, the Swiss theologian Eduard Thurneysser. Thurneysser's study of Dostoevsky (1921) served as a pivotal stimulus for Barth, revealing to him the depth of the Russian writer's anthropological and existential insights. In this case, reception consisted not so much in assimilating the literary form as in absorbing the fundamental philosophical categories. Central for Dostoevsky, as highlighted by Thurneysser, was the question "What is man?" (Rae, 1970: 77). His characters, who remain human rather than divine and whose appeals resemble the evangelical, became for Barth the point of departure in his critique of humanistic optimism. The Christian response found in Dostoevsky — the idea of "resurrecting life" as the salvation granted

by God — was incorporated into the foundations of "crisis theology." This reception of fundamental ideas concerning man, sin, and salvation allowed Barth to significantly develop his early theology (Rae, 1970: 77).

Thus, both cases demonstrate that the phenomenon of reception is a process of active transformation. Whether linguistic adaptation to conform to aesthetic norms or through philosophical interpretation for integration into a new theological system, the source work inevitably passes through the formative filter of a mediator. As a result, the receiving culture acquires not the original in its pure form but its adapted version, which nonetheless proves capable of exerting a profound and at times revolutionary impact on new intellectual soil — a phenomenon that undoubtedly merits dedicated, in-depth study.

The creation of the first complete translation of *Crime and Punishment* in 1886 coincided with key social processes. The Meiji period became a time of radical restructuring of Japanese society, where the collision of traditional values with Western influences generated a unique synthesis. Urbanization, driven by industrialization, led to the mass migration of the rural population to cities. Tokyo, Osaka, and Nagoya were transformed into centers of the new economy, where factories, banks, and educational institutions were concentrated (Minami, 1967: 1, 8, 9, 18). However, this process was accompanied by sharp property stratification: the peasantry, dispossessed of land as a result of the 1873 reforms, swelled the ranks of the urban proletariat, while the samurai elite and the emerging bourgeoisie accumulated capital. Social tension was exacerbated by the contrast between the luxury of the new Ginza districts and the slums of Asakusa, where poverty and disease prevailed. These realities were reflected in the journalism of the time: naturalist writers such as Kunikida Doppo depicted the fates of the "lost generation," torn from the patriarchal order and cast into the vortex of capitalist relations (Brecher, 2012: 5).

Against this backdrop, an ideological conflict unfolded between the Buddhist heritage and Christian missionaries. After the ban on Christianity was lifted in 1873, Protestant preachers poured into Japan, and their activities were perceived as a threat to traditional institutions. Buddhist schools, having lost state support following the separation of Shinto in 1868, responded with a campaign for the "purification of doctrine" — the *hanshukyō undō* movement (Grapard, 1984: 241). The polemics between Buddhist monks and Christians were conducted not only in temples but also in the pages of newspapers: in the 1880s, debates regarding the nature of suffering — where Christians insisted on the redemptive sacrifice of Christ and Buddhists on

overcoming *dukkha* through the Eightfold Path — became part of public life. Paradoxically, this conflict stimulated a philosophical synthesis: thinkers such as Inazo Nitobe reinterpreted Buddhist concepts through the lens of Western humanism, laying the groundwork for the later "Kyoto School" philosophy (Stone, 2021: 3; K. Hung, 2009: 242).

At the intersection of Japanese traditions and the active integration of Western thought, a new ethics emerged as a response to the challenges of modernization. Confucian principles, which had long regulated social relations, adapted to the realities of industrial society. The ideal of *jin* (benevolence) transformed into the concept of civic duty, as reflected in the elementary school textbooks *Shōgaku Shushinshō*, where loyalty to the emperor was integrated with technical education (Shimbori, 1960: 98). Concurrently, Western rationalism, disseminated by the Meirokusha society, introduced notions of utilitarianism and individual rights (Ghadimi, 2017: 207; Lin & Lu, 2019: 40).

Philosopher Yukichi Fukuzawa, in his essay "An Encouragement of Learning" (1872–1876), asserted that the Confucian virtue of *gi* (righteousness) should be combined with *benri* (practical utility) (Cheng, 2013: 22). This synthesis engendered a unique ethical system wherein the collectivism of traditional morality coexisted with individualistic aspirations, manifesting most vividly in the women's rights movement and the family law reforms of 1898 (Takakusu, 1906: 7).

Thus, the Meiji era became a springboard for cultural synthesis, where social upheavals, ideological conflicts, and ethical quests formed the foundation of modern Japanese identity. The interaction of these factors not only determined the paths of modernization in the early Meiji period but also created the groundwork a unique interpretation of Western texts, in which traditional categories served as a bridge between civilizations. A significant contribution to the reception of Dostoevsky in Japan was made by the translator Masao Yonekawa, whose translation of *Crime and Punishment* was first published in 1935 and later revised for the complete collected works of the writer (completed in 1953 by Kawade Shobo publishing house). This translation emerged within a cultural context of interest in Russian literature, which reflected the pursuit of profound psychological and ethical analysis amid social upheavals and the intellectual synthesis of East and West. Yonekawa, a graduate of the Tokyo School of Foreign Languages, contributed to the popularization of Dostoevsky, for which he received the Yomiuri Literary Prize, thereby underscoring the evolution of this reception from the Meiji era to the present day. Within the framework of

this study, it is precisely his translation that will be examined, as the most thoroughly elaborated and meticulously refined.

The study of semantic transformations in F. M. Dostoevsky's novel *Crime and Punishment* in the Japanese translation of the Meiji period opens a unique window into the process of mutual penetration of philosophical systems. The object of analysis is not linguistic equivalence but the ontological recoding of meanings — from Christian providentialism to the Buddhist doctrine of emptiness, from Russian existential reflection to Confucian ethics of duty. Each semantic cluster — be it "suffering," "freedom," or "fate" — appears as a node in a complex network of cultural correspondences where the translator acts as a mediator between civilizational codes.

The relevance of the translation study lies in its ability to deconstruct the very mechanism of intercultural communication in the philosophical context. In the era of globalization, when the "East-West" dialogue is often reduced to the superficial borrowing of forms, the analysis of Meiji translation strategies demonstrates an alternative model — a deep synthesis in which a foreign idea gains new life in a different philosophical soil. Thus, the concept of "conscience," for example, rooted in Christian metaphysics of sin, is reborn on Japanese soil as 良心 (*ryōshin*) — an innate Confucian virtue, preserving its ethical charge but changing its ontological foundation.

The methodological framework of the study combines corpus analysis with the principles of comparative philosophy. The calculation of semantic density in clusters enables a quantitative evaluation of cultural priorities: if existential reflection dominates in the original, the translation accentuates a sense of social duty. However, behind these figures lies a qualitative shift: the Russian "crime" as a transgression of divine law transforms into 罪 (*tsumi*) — a notion integrated into both the Buddhist concept of karma and Confucian views on social harmony.

The philosophical significance of our study resides in demonstrating how the translation practices of the Meiji era generated space for the emergence of novel meanings. Russian existentialism, upon encountering the Zen concept of 無 (*mu* — non-being), gives rise to hybrid forms: Raskolnikov's "despair" is reconceptualized through 諦め (*akirame* — resignation), which merges stoicism with the notion of non-attachment. This synthesis was not a matter of mechanical appropriation — it emerged as a response to the pressure of modernization, wherein traditional values required reformulation in the lexicon of the emergening epoch.

The historical and cultural context of the study reveals the Meiji paradox: the drive toward westernization achieved through a return to tradition.

The translation of *Crime and Punishment* constituted an integral component of the endeavor to construct an "enlightened nation," wherein Western ideas were filtred through the lens of localization.

The transformation of the "Fate" cluster reflects this duality: Christian fatalism is supplanted by the Buddhist engi (interdependent origination), yet simultaneously imbued with the Confucian pathos of social responsibility. The study contributes to the philosophy of language by demonstrating that translation is not merely the transmission of information but a creative act of constructing a new philosophical reality. Through the analysis of semantic shifts, it becomes evident that Russian literature, filtered through the prism of Japanese thought, acquires the qualities of a cultural archetype — universal and local in equal measure. This process transcends the bounds of literary studies, offering a model for comprehending contemporary inter-civilizational interactions, wherein dialogue occurs not at the level of borrowings but through profound transformation of meanings.

The study of cultural transformations in the translation of literary text necessitates an approach that overcomes the quality-quantity dichotomy. The article works with concepts such as "semantic density" and "cultural index," which, as formal metrics, acquire heuristic value only within the context of philosophical reflection on the nature of cultural transfer. These notions, at first glance belonging to corpus linguistics, emerge as instruments for deconstructing profound anthropological structures, revealing how linguistic practices shape the ontological horizons of human existence.

The concept of semantic density (SD), calculated as the ratio of the frequency of lexical units in the cluster to the overall volume of the text, servers as an indicator of cultural perception (Ge, 2022: 6, 12, 13). In the context of comparative analysis between the original and the translation, this metric enables the detection of implicit strategies of cultural adaptation, where the quantitative dominance of certain semantic fields marks zones of heightened relevance. This method overcomes the limitations of purely qualitative analysis, offering a verifiable foundation for comparing cultural axiological systems. In the philosophy of culture, this metric takes on the status of a "cognitive magnet" (by analogy with E. Rosch's theory),[1] where the concentration of certain concepts marks zones of semantic tension. For instance,

[1]In Rosch's theory, prototypes function as reference instances of a category endowed with maximum representativeness (e. g., "sparrow" for the category "birds") (Rosch, 1975: 2, 3, 34). This mechanism carries profound implications for the philosophy of language: lexical units with elevated semantic density, discerned in the analysis of the translation of *Crime and Punishment*, operate as prototypes/magnets, structuring the text's perception via culture-specific filters.

the increase in the density of the "Fate" cluster in the Japanese translation does not merely reflect a statistical anomaly but signals a fundamental shift in the understanding of temporality: Christian providentialism, presupposing linear progression toward the eschaton, is replaced by the Buddhist concept of samsara with its cyclical model of time. In the context of philosophical anthropology, this phenomenon can be interpreted through the prism of M. M. Bakhtin's theory of the "chronotope," wherein the spatio-temporal coordinates of the text determine the anthropological model (Bakhtin, 1975). The heightened frequency of 縁起 (*engi* — interdependent origination) in place of "fate" transforms the very image of the human: from a subject who challenges the transcendent order (Raskolnikov), the character becomes an element of a karmic network, where individual choice dissolves into a chain of causal connections.

Furthermore, this article seeks to introduce the concept of "cultural index" (CI), which enables a quantitative assessment of the degree of conceptual adaptation of the text to the value-semantic matrices of the target culture. Calculated as the ratio of the semantic density (SD) of the translation to the SD of the original, this indicator serves as a measure of cultural relevance for thematic clusters, revealing zones of heightened attention or deliberate reduction. The CI fulfills a dual function: on the one hand, it registers statistical anomalies (deviations from the source semantic structure); on the other, it acts as a hermeneutic key for interpreting cultural filters (Chernikova et al., 2020). Thus, the decline in the index for the psychological cluster ($CI = 0.79$, below unity) correlates with the Buddhist negation of a persistent "self," minimizing interest in individualized introspection. For example, the increase in mentions of 良心 (*ryōshin* — conscience) is accompanied by a semantic shift: from the Christian "inner voice" ascending to the Augustinian notion of the divine spark in humanity to its reconceptualization as the Confucian "innate virtue" according to Mencius (Jiang, 1997: 269), fundamentally altering the anthropological model. The analysis by V. S. Stepin of the dynamics of cultural transmissions allows the interpretation of translation as an ontologically creative act, wherein the reconstruction of meanings generates novel epistemological realities (Stepin, 2006). The cultural index becomes a measure of this creative transformation, where quantitative change represents the surface manifestation of profound semiotic processes.

For instance, the hypertrophy of the term「義理」(*giri*) in the Japanese version emerges as a cognitive magnet, redirecting the semantics of ethical choice toward the realm of social duty.

The application of mathematical methods in humanities research ensures the objectification of cultural patterns, revealing latent semantic shifts through quantitative analysis of frequency and the cultural index, which, in conjunction with hermeneutic interpretation, allows for overcoming the subjectivity of qualitative approaches while preserving the depth of philosophical-anthropological analysis of meaning transformations in intercultural space. These methods, in the humanities context, are often criticized for reductionism; however, in this framework, quantitative indicators render visible those cultural patterns that remain concealed in purely qualitative analysis. The frequency of verbs of motion in *Crime and Punishment*, for instance, does not merely indicate stylistic preferences, but unveils a fundamental divergence in the understanding of human activity: in the original, terms like "went," "stood up," and "sat down" mark discrete actions of the subject asserting its will, whereas in the Japanese translation, 歩み (*ayumi* — movement/path) and 待つ (*matsu* — waiting) emphasize the processuality of being, aligning with Zen philosophy's doctrine of spontaneous nature. This contrast can be interpreted through the opposition of "agency versus processuality" proposed by anthropologist T. Ingold (Ingold, 2006: 10). The Russian text embodies the Western model of the subject as the source of actions, while the Japanese translation represents the Eastern conception of the human as a participant in the universal flow of changes.

The methodology applied in the article draws upon the concept of "cultural filters" by Y. M. Lotman, according to which translation constitutes not the transmission of information, but its recoding through a system of cultural codes (Lotman, 2000: 117). Lotman's notion of cultural filters is essential for interpreting translation as a semiotic act of recoding, wherein the transformation of meanings is conditioned by the interaction of discursive systems from the source and target cultures, enabling the analysis of cultural adaptation mechanisms through the prism of contextual codes. The semantic shift analysis employed in the study of the novel *Crime and Punishment* translation, particularly in the "Poverty" cluster, demonstrates how the social issues of the original are filtered through the Confucian principle 修身斉家治国平天下 (*shūshin seika chikoku heitenka* — cultivating oneself to bring peace to the world), Buddhist compassion, and the social Darwinist ideas of the Meiji era: the interpretation of Darwin's theory through the lens of Confucian ethics and Buddhist ontology engenders a unique conception of progress, wherein "natural selection" was regarded not as biological fate, but as a moral imperative for national self-perfection. In the context of the work's analysis, where the "Poverty" cluster plays a significant role, it is

worth noting that this conception manifested in both urbanization and pauperization: industrial growth led to the formation of a working class whose condition was justified by the theory of "natural selection." Poverty was construed as a consequence of personal inability to adapt, which is reflected in the absence of systemic social support until 1911 (Taira, 1969: 165).

The concept of 八紘一宇 (*hakkō ichiu* — "the eight corners of the world under one roof"), emerging in the early twentieth century, represented an ultranationalist doctrine proclaiming Japan's mission to unite the world under the "single roof" of imperial authority, thereby legitimizing colonial expansion through the rhetoric of "civilizing duty" and racial superiority. This concept employed social Darwinist rhetoric to justify the annexation of Korea (1910) as a "civilizing mission," while the system of competitive examinations for officials (introduced in 1887) was interpreted as a mechanism of "natural selection" for the finest minds, although in practice it reproduced samurai hierarchies (Nirei, 2011).

This synthesis of social Darwinism with Confucian and Buddhist ideas engenders a unique hybrid: Raskolnikov's "beggar" becomes 非人 (*hinjin* — non-person, an outcast beyond the caste system), linking medieval Japanese marginality with images of urban representatives of the lower strata of the modern era. The cultural index here serves as a measure of the intensity of cultural projection, specifically the translator's capacity to appropriate foreign social experience through local categories. This process illustrates how translational activity, functioning as an existential practice, unveils a fundamental divergence between Western subjectivity — centered on the reflexive "I" (Descartes, Kierkegaard (Pörn, 1984)) — and Eastern anthropology of 無我 (*muga* — "no-self"), rooted in the Buddhist denial of *ātman* (Andersen, 2020: 38), wherein analysis of the "Psychology" cluster reveals a reduction of individualized experience in favor of collectivized reality. The semantic analysis methods employed in the study (semantic density, cultural index, comparative hermeneutics) thus transform their instrumental role, becoming a form of philosophical reflection on the nature of cultural interaction. Semantic density and the cultural index emerge not merely as metrics, but as concepts illuminating the dialectics of preservation and change in the process of inter-civilizational dialogue. Through their lens, translation appears not as a technical operation, but as an anthropological act — a space wherein a new image of humanity is born, synthesizing ostensibly incompatible cultural codes. This methodological perspective opens avenues for reconceptualizing the very nature of cultural identity in the era of globalization, where traditional oppositions of "East-West" yield to complex hybrid forms.

Upon turning to Dostoevsky's original text, we observe that the psychological depth of characters unfolds through an inner dialogue with the absurdity of existence, wherein feelings of guilt, fear, and despair serve as markers of existential crisis. However, in the Japanese version, these emotional states are reconceptualized through the prism of the Buddhist ontology of "no-self." The doctrine of *anātmavāda*, constituting the core of Buddhist ontology, negates the existence of a permanent substantial "self" (*ātman*), viewing human personality as a transient aggregation of the five *skandhas* (form, sensation, perception, mental formations, and consciousness) (Chadha, 2017: 1). This principle radically transforms the understanding of emotional states: whereas in the Western tradition "remorse" presupposes a stable subject bearing responsibility for its actions, the Japanese 後悔 (*kokai*) accentuates situational regret over disruption of social order. Such an interpretation stems from the Buddhist conception of "no-self" — the emotion does not belong to the individual but arises as a product of the *skandhas'* interaction in specific circumstances (Gallagher et al., 2023: 1). In the context of translation, this leads to a diminution of existential reflection: Raskolnikov's inner dialogue with his "self" is supplanted by an examination of the external consequences of the act.

The aforementioned shift in emphasis from individual reflection to collective responsibility mirrors the Confucian ideal of 和 (*wa* — harmony), wherein personal experience is subordinated to the maintenance of group equilibrium (Feng & Newton, 2012: 342). The Confucian conception of wa, rooted in the Lunyu (Analects (The Analects of Confucius, Watson, 2007)), posits harmony as the foundational principle of cosmic and social order (Cheng, 2006: 26). In contrast to the Western accent on individualism, *wa* underscores the interdependence of all elements within the system (Kim et al., 2010). This manifests in the translational strategy through the socialization of ethics, as "conscience" — conceived as an inner voice (transforms into 良心 (*ryōshin*)) — an innate virtue oriented toward sustaining group solidarity; and the collectivization of emotions, as "shame" (shame before oneself) becomes 恥 (*haji*) — shame before society, aligning with the Confucian maxim: "The noble man feels shame when his words diverge from his deeds" (Lebra, 1983: 205).

Even "suffering," central to the existential narrative, is recoded as 苦 (*ku*) — the foundational category of the Four Noble Truths, converting personal drama into an illustration of the universal law of *dukkha* (Gäb, 2015: 346). The first noble truth of Buddhism — "all is suffering" (*dukkha*) — finds paradoxical embodiment in the translation. Whereas in Dostoevsky, characters'

suffering bears an existential character (conflict with the transcendent), the Japanese 苦 (*ku*) accentuates its universality and inherent naturalness. This displacement accords with the theory of 諸行無常 (*shōgyō mujō* — impermanence of all phenomena), as suffering appears in the translation as a consequence of attachment — the hypertrophy of the term (24 instances in the original versus 125 mentions in the translation) underscores the Buddhist notion that *dukkha* arises from the desire to retain the impermanent. Such an approach reflects not only Buddhist ontology but also its ethico-therapeutic imperatives, wherein mental health is understood as a consequence of moral virtue, and psychic disharmony as the outcome of a "disordered" character incapable of maintaining thoughts and feelings in proper order (Balogh, 2020: 125). The translation likewise eschews individualized descriptions, presenting suffering as the common lot of samsaric existence, rendering the experience not personal but collective (Gäb, 2015: 349). Thus, the hero's psychological crisis is interpreted not as an existential revolt, but as a moral-ontological delusion requiring rectification through the acceptance of universal Buddhist truths.

Regarding the novel's moral dilemmas, they undergo a radical transformation within the Japanese cultural context. The Christian "conscience" as the voice of transcendent truth is supplanted by 良心 (*ryōshin*) — an innate virtue rooted in the Confucian system of 五倫 (*gorin* — the five relationships). In the original *Crime and Punishment*, Raskolnikov's conscience functions as evidence of divine presence (referencing the Augustinian concept of the "inner teacher," wherein conscience constitutes the spark of divine reason within humanity (Svensson, 2012: 4)), an instrument of existential choice (the pangs of conscience following the old woman's murder represent not merely emotion, but an ontological rupture with the divine order), and individual responsibility (the character stands alone before eternity, aligning with the Protestant ethic of *solus cum Deo* — alone with God). This model derives from the biblical tradition: "For when Gentiles... the work of the law is written in their hearts, about which their conscience bears witness." In the novel, the frequency of the word "conscience" (24 instances) correlates with the character's crisis as a metaphysical rebel. The Japanese term 良心 (*ryōshin*), literally "good heart," reconceptualizes the notion through the Confucian lens of innate virtue as per Mencius;[2] a socially conditioned

[2]"Human nature is good, just as water flows downwards" (The Chinese Classics..., Legge, 1869: 59).

ethics wherein conscience is directed not toward dialogue with the transcendent, but toward fulfilling one's role in the hierarchy — for instance, as a son, subject, or friend — and ritualized behavior.[3] This translational choice acquires particular significance amid the intellectual debates of the Meiji era. The reinforcement of Confucian ethics of duty clashed with another influential current — "Buddhist modernism," which, conversely, sought to synthesize Buddhism with ideals of the European Enlightenment, such as individualism, universalism, and personal freedom (Shields, 2022: 319). Thus, the novel's translation emerges not merely as an act of linguistic adaptation but as a deliberate ideological gesture wherein preference was accorded to the preservation of collectivist Buddhist morality in the face of modernist European challenges extolling individual autonomy.

The presented Japanese translation transforms the ethical landscape of the work: if Raskolnikov rebels against the divine order in Dostoevsky, his Japanese counterpart disrupts the "natural harmony" (自然の調和), necessitating restoration through a ritual of atonement within the social hierarchy. Even the concept of "crime" acquires new dimensions: 罪 (*tsumi*) in the translation bears the nuance of karmic imbalance, demanding not punishment but the restoration of 理 (*ri* — cosmic order). Thus, for example, the Japanese text post-translation envisions when Raskolnikov reflects on the "right to crime," the Japanese translation employs 義理 (*giri* — social duty), linking the moral conflict to a violation of horizontal relationships rather than the vertical "man-God."

The novel's temporal structure — originally built by Dostoevsky on the tension between past crime and future punishment — dissolves in the Japanese rendition into the Buddhist concept of 無常 (*mujō*, impermanence). Verbs of motion that mark discrete acts of will in the original ("went," "stood up") are replaced by processual forms such as 歩み (*ayumi*, "path") and 続く (*tsuzuku*, "continuation"), emphasizing the continuity of being. This shift reflects a fundamental divergence in conceptions of human activity: the Western subject as the source of action encounters the Eastern idea of 無為自然 (*muyi shizen*, "spontaneous following of the flow of reality"). In Dostoevsky, time serves as an arena of metaphysical confrontation — discrete markers denote crucial moments of existential choice, and temporal indicators ("yesterday," "minute," "hour") heighten the tension between

[3] In the translation, "conscience" appears 22 times, but the context is shifted towards the Confucian 礼 (*rei* — ritual propriety).

the irreversible past and the eschatological future. The Japanese translation, by contrast, foregrounds being's processuality through nominalization: substituting 歩み (*ayumi*, "path") for "went" shifts attention from a volitional act to continuous movement in line with the Zen concept of 道 (*dō*), where the goal dissolves in the process; the hypertrophy of unfinished verb forms (待ち続けた, *machi tsuzuketa*, "continued waiting" instead of "waited") introduces a Buddhist view of time as meditative anticipation — pauses are integral to action; and cyclical reference (replacing "yesterday" with 一度 (*ichido*, "one time")) actualizes 縁起 (*engi*, interdependent origination), viewing events not as unique points but as links in a rebirth chain. Here, Raskolnikov ceases to be the autonomous agent of his deeds and becomes a "conduit" for karmic processes. His "crime" no longer appears as a volitional act but as 業 (*gō*, the inevitable result of past actions). Even the murder is rendered as 斬り続けた (*kiri tsuzuketa*, "continued chopping"), erasing the boundary between act and consequence and relocating moral judgment to the restoration of 理 (*ri*, cosmic balance). Urbanization likewise assumes a temporal dimension: references to 流れ (*nagare*, "flow") in the context of city life (46 occurrences) mirror Meiji modernization, when time was conceived as a "river" carrying the individual beyond the bounds of personal volition.

Thus, the Meiji translators, navigating between Westernization and tradition, forged a chronotopic hybrid of linear progress time intertwined with a cyclical historical perception through the prism of 王朝循環論 (*ōchō junkanron* — dynastic cycle theory) (Moniz Bandeira, 2020: 2) and the Confucian ritual of 礼 (*rei*), which structured everyday life, transforming the chaos of modernization into an ordered flow of 序 (*jo* — sequence).

The adaptation of the "poverty" cluster unveils an anthropological opposition between individualism and collectivism. The Russian "beggar," as a symbol of existential solitude, is supplanted by 非人 (*hinin*) — a historical term for the caste of outcasts from the Edo era. This transformation redirects the emphasis from personal tragedy to systemic inequality, legitimated by the Confucian principle of 義 (*gi* — social righteousness) (Chen, 2020: 2). Poverty ceases to represent a spiritual condition, emerging instead as a marker of caste affiliation, which mirrors the realities of the Meiji period when urbanization exacerbated contradictions between feudal remnants and capitalist relations (Taira, 1969: 156). The phenomenon of *hinin* — the caste of "non-humans" during the Edo period (1603–1868) — embodies a paradoxical realization of the Confucian principle of *gi*, wherein social marginalization was justified as an essential precondition for upholding harmonious order (Amos, 2017: 581). Hinin, engaged in "unclean" occupations

(slaughtering animals, disposing of corpses, executing sentences), existed outside the four-tier class system (*shi-nō-kō-shō*), becoming the living embodiment of the "other" in the Confucian societal model (Smythe, 1952: 194). Their status was not the result of personal failings but predetermined by birth, aligning with the Confucian notion of 分 (*bun*) — the immutable division of social roles (Nuyen, 2001: 62). In the Japanese translation of *Crime and Punishment*, the substitution of "beggar" with 非人 (*hinin*) actualizes not an individual tragedy but a systemic hierarchy, wherein poverty signifies not a spiritual state but a caste marker. This reflects the essence of *gi* as "justice through differentiation": *hinin*, akin to Dostoevsky's characters, prove necessary for demarcating the boundaries of "normal" society. Their existence was legitimated through the Confucian maxim 君君臣臣父父子子 (*kun kun shin shin fu fu shi shi* — "let the ruler be ruler, the subject subject, the father father, the son son"), wherein each element of the system acquires meaning through opposition to the "other" (Guo, 2013: 62). Yet this logic clashed with the Buddhist principle of 平等 (*byōdō* — universal equality), generating tension in the perception of *hinin* (J. Hung, 2020: 312). Meiji-era translators, seeking to adapt Dostoevsky's social critique, employed this term as a bridge between Christian compassion for the downtrodden and the Japanese concept of 慈悲 (*jihi* — compassion), oriented not toward systemic change but toward the ritual "purification" of suffering via acceptance of karmic predetermination. Thus, the semantic shift from "beggar" to 非人 (*hinin*) in the translation constitutes an act of cultural hermeneutics, wherein the Confucian ideal functions not as an ethical imperative but as an instrument for conserving social ontology, with hinin serving as the living embodiment of the boundary between the "human" and the "non-human" in a hierarchized world.

The concept of fate undergoes the most profound ontological shift. Christian fatalism, presupposing linear progression toward an eschatological finale, dissolves into the Buddhist model of 縁起 (*engi* — interdependent origination). Raskolnikov's death, in the original acquiring meaning through redemption, is reinterpreted as 成仏 (*jōbutsu*) — the completion of the rebirth cycle. This transformation alters the very image of humanity: from a contender against the transcendent order, the figure emerges as a traveler awakening to his role in the samsaric cycle. Engi, a cornerstone of Buddhist ontology, denotes the principle of the interdependent arising of all phenomena (Kardash·ch, 2015: 293). In the context of the *Crime and Punishment* translation, this concept radically reconceptualizes understandings of fate and responsibility. Whereas in the original, Raskolnikov's fate is framed through Christian

providence (linear trajectory toward redemption), the Japanese version introduces 因果応報 (*inga ōhō* — karmic retribution), wherein each action constitutes a link in an infinite chain of cause and effect. This transition converts the existential drama into a narrative of restoring disrupted moral order, consonant with traditional Eastern views on the inextricable bond between morality and well-being (Balogh, 2020: 125). For instance, "fate" appears only 8 times in the original text, contrasted with 37 instances of engi. The shift accentuates not predetermination but the dynamic interrelation of actions: Raskolnikov's crime is no longer a challenge to divine order but a disruption of 理 (*ri* — cosmic balance), demanding restoration through the chain of rebirths. The term *jōbutsu* (成仏) — "attaining Buddhahood" — precisely signifies the termination of the samsaric cycle via liberation from passions. In the translation, this notion reconceives the novel's denouement, as evidenced by the semantic shift: "resurrection" is mentioned 4 times in the original, while *jōbutsu* appears only once yet bears conceptual weight. The hero's death is construed not as physical cessation or Christian soul resurrection but as transition into the state of 涅槃 (*nehan* — nirvana), where suffering is transcended through dissolution into "emptiness." The emphasis on traditional doctrines of karma and rebirth may also be viewed as an ideological choice by the translators. In an era when Buddhist modernists advocated revising "religious orthodoxy" in favor of more rational and scientific paradigms (Shields, 2022: 319), the translators of *Crime and Punishment* leveraged Dostoevsky's text to affirm and revitalize the foundational tenets of the traditional Buddhist-Confucian worldview. The cultural-philosophical context of these concepts within the study underscores the importance of accounting for modernization and inter-confessional dialogue amid the socio-cultural transformations of the Meiji period. During Westernization, the notion of interdependence facilitated the synthesis of Buddhist tradition with scientific determinism (McMahan, 2004: 900). The sociologist Inazō Nitobe, in *Bushidō* (1899), likened karma to the "natural laws" of Western science (Nitobe, 1914: 117). In the translation, this manifests through the hypertrophy of the term 因果 (*inga* — cause and effect) — from 0 in the original to 43 mentions. The absence in Buddhism of a Last Judgment concept led to the replacement of "redemption" with 解脱 (*gedatsu* — liberation). Raskolnikov's scene of repentance is depicted via 悟り (*satori* — enlightenment), wherein guilt is overcome not through suffering but through realization of the "self's" illusoriness. Thus, *engi* negates the autonomous agent — Raskolnikov becomes a "conduit" for karmic processes rather than the author of his crime — while

jōbutsu redirects focus from personal salvation to dissolution in the absolute, reflecting Buddhism's critique of attachment to the "self."

The features we have identified, we contend, transform the Japanese translation of *Crime and Punishment* into more than a mere linguistic artifact; it emerges as a philosophical endeavor in reconceptualizing time — a domain wherein Christian existentialism intersects with Eastern ontology of process, thereby generating novel horizons for comprehending human existence amid the epoch of global transformations.

The investigation of semantic transformations in the Japanese translation of F. M. Dostoevsky's novel *Crime and Punishment* unveils a fundamental paradox of intercultural communication: the more faithfully the translation replicates the text's surface structure, the more radically it reconstitutes its ontological foundations. Each semantic shift — whether the substitution of "conscience" with 良心 (*ryōshin*) or the reconceptualization of "fate" through 縁起 (*engi*) — constitutes an act of philosophical creativity, wherein Christian existentialism, Buddhist ontology of process, and Confucian ethics of duty collide and mutually enrich one another. In our view, a pivotal outcome warranting emphasis is the delineation of two interconnected levels of cultural transfer. At the semiotic level, lexical substitutions activate the deep structures of the collective unconscious — archetypes such as「和」(*wa*, harmony) and「空」(*kū*, emptiness) — that shape Japanese perception of reality. At the ontological level, semantic clusters reconfigure the very "substance" of the narrative, converting a linear drama of individualized choice into a cyclical parable of karmic equilibrium. These transformations expose a principled divergence in the constitution of the human subject: whereas Dostoevsky's original embodies the tragedy of the "I" challenging the transcendent, the translation delineates a portrait of the "non-I," dissolved within a web of social and karmic interconnections. Moral dilemmas, initially rooted in the notion of freedom, are reformulated through the Confucian principle of 義 (*gi*), wherein ethics functions not as an internal imperative but as a mechanism for sustaining cosmic order. We posit that this phenomenon of Meiji-era translation manifests as a process in which Western ideas, filtered through the prism of traditional Japanese thought, acquire new semantic flesh. This was neither imitation nor distortion but a form of cultural appropriation, wherein the "foreign" served as a catalyst for reinterpreting the "native." Thus, it becomes evident that translation constitutes a full-fledged philosophical practice, with language functioning as the medium for birthing hybrid ontologies. From the vantage of philosophical anthropology, the findings corroborate the thesis of the inherently cultural conditioning

of human experience. The Russian existential revolt and the Japanese acceptance of 理 (*ri*) emerge not as antitheses but as divergent modalities for apprehending a singular archetypal narrative — the encounter of humanity with the limits of its own freedom. Translation, accordingly, becomes a space wherein these modalities engage in dialogue, engendering fresh horizons for understanding what it means "to be human" in a globalizing world.

REFERENCES

Amos, T. D. 2017. "The Subaltern Subject and Early Modern Taxonomies: Indianisation and Racialisation of the Japanese Outcaste." *Asian Studies Review* 41 (4): 577–593.

Andersen, M. B. 2020. "Identity and the Elusive Self: Western and Eastern Approaches to Being No One." *Journal of Sport Psychology in Action* 11 (4): 243–253.

Bakhtin, M. M. 1975. *Voprosy literatury i estetiki [Questions of Literature and Aesthetics]* [in Russian]. Moskva [Moscow]: Khudozhestvennaya literatura.

Balogh, L. 2020. *Psychotherapy in East Asia: A Philosophical and Historical Perspective*. Cham: Springer Nature Switzerland AG.

Brecher, W. P. 2012. "Useless Losers: Marginality and Modernization in Early Meiji Japan." *The European Legacy* 17 (6): 803–817.

Chadha, M. 2017. "No-Self and the Phenomenology of Ownership." *Australasian Journal of Philosophy* 96 (1): 14–27.

Chen, S. 2020. "The Official Discourse of Social Justice in Citizenship Education: A Comparison between Japan and China." *Education, Citizenship and Social Justice* 16 (3): 197–210.

Cheng, C. 2006. "Toward Constructing a Dialectics Of Harmonization: Harmony And Conflict in Chinese Philosophy." *Journal of Chinese Philosophy* 33 (S1): 25–59.

——— . 2013. "Confucian Ethics in Modernity: Ontologically Rooted, Internationally Responsive, and Integratively Systematic." *Journal of Chinese Philosophy* 40:76–98.

Chernikova, N. V., I. V. Sidorova, and V. M. Shvetsova. 2020. "Linguo-Conceptual Analysis as an Effective Technology for Organizing Scientific and Educational Activities." *Journal of Physics: Conference Series* 1691 (1).

Feng, L., and D. Newton. 2012. "Some Implications for Moral Education of the Confucian Principle of Harmony: Learning from Sustainability Education Practice in China." *Journal of Moral Education* 41 (3): 341–351.

Gäb, S. 2015. "Why Do We Suffer? Buddhism and the Problem of Evil." *Philosophy Compass* 10 (5): 345–353.

Gallagher, S., A. Raffone, A. Berkovich-Ohana, et al. 2023. "The Self-Pattern and Buddhist Psychology." *Mindfulness* 15 (4): 795–803.

Gavin, M. 2004. "National Moral Education: Abe Isô's Views on Education." *Japanese Studies* 24 (3): 323–333.

Ge, Y. 2022. "The Linguocultural Concept Based on Word Frequency: Correlation, Differentiation, and Cross-cultural Comparison." *Interdisciplinary Science Reviews* 47 (1): 3–17.

Ghadimi, A. 2017. "The Federalist Papers of Ueki Emori: Liberalism and Empire in the Japanese Enlightenment." *Global Intellectual History* 2 (2): 196–229.

Grapard, A. G. 1984. "Japan's Ignored Cultural Revolution: The Separation of Shinto and Buddhist Divinities in Meiji ('Shimbutsu Bunri') and a Case Study: Tōnomine." *History of Religions* 23 (3): 240–265.

Guo, Q. 2013. "On Confucian Political Philosophy and Its Theory of Justice." *Frontiers of Philosophy in China* 8 (1): 53–64.

Hung, J. 2020. "Is Dharma-Nature Identical to Ignorance?: A Study of 'ji' in Early Tiantai Buddhism." *Asian Philosophy* 30 (4): 307–323.

Hung, K. 2009. "Alien Science, Indigenous Thought and Foreign Religion: Reconsidering the Reception of Darwinism in Japan." *Intellectual History Review* 19 (2): 231–250.

Ingold, T. 2006. "Rethinking the Animate, Re-animating Thought." *Ethnos* 71 (1): 9–20.

Jiang, X. 1997. "Mencius On Human Nature And Courage." *Journal of Chinese Philosophy* 24 (3): 265–289.

Kardash·ch, G. 2015. "From Etymology to Ontology: Vasubandhu and Candrakīrti on Various Interpretations of Pratītyasamutpāda." *Asian Philosophy* 25 (3): 293–317.

Kim, J., T. Lim, K. Dindia, and N. Burrell. 2010. "Reframing the Cultural Differences between the East and the West." *Communication Studies* 61 (5): 543–566.

Lebra, T. S. 1983. "Shame and Guilt: A Psycho Cultural View of the Japanese Self." *Ethos* 11 (3): 192–210.

Lin, Z., and H. Lu. 2019. "In Search of a Moral Standard: Debates over Ethics Education and Religion in Meiji Japan." *History of Education* 49 (1): 38–56.

Lotman, Yu. M. 2000. *Semiosfera [Semiosphere]* [in Russian]. Sankt-Peterburg [Saint Petersburg]: Iskusstvo-SPb.

McMahan, D. L. 2004. "Modernity and the Early Discourse of Scientific Buddhism." *Journal of the American Academy of Religion* 72 (4): 897–933.

Minami, R. 1967. "Population Migration Away from Agriculture in Japan." *Economic Development and Cultural Change* 15 (2): 183–201.

Moniz Bandeira, E. 2020. "From Dynastic Cycle to Eternal Dynasty: The Japanese Notion of Unbroken Lineage in Chinese and Korean Constitutionalist Debates, 1890–1911." *Global Intellectual History* 7 (3): 517–532.

Nirei, Y. 2011. "Globalism and Liberal Expansionism in Meiji Protestant Discourse." *Social Science Japan Journal* 15 (1): 75–92.

Nitobe, I. 1914. *Bushido, the Soul of Japan.* 20th ed. Kojimachi and Tokyo: Teibi Publishing Company.

Nuyen, A. T. 2001. "Confucianism and the Idea of Equality." *Asian Philosophy* 11 (2): 61–71.

Pörn, I. 1984. "Kierkegaard and the Study of the Self." *Inquiry* 27 (1–4): 199–205.
Rae, S. H. 1970. "Dostoevsky and the Theological Revolution in the West." *The Russian Review* 29 (1): 74–80.
Rosch, E. 1975. "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology* 104 (3): 192–233.
Shields, J. M. 2022. "Zen Internationalism: Inoue Shûten, Uchiyama Gudô, and the Crisis of (Zen) Buddhist Modernity in Late Meiji Japan." In *Waves of Radicalism : Global Politics in the Tides of Revolution*, ed. by C. Tudor and K. Kornetis, 319–344. Lanham: Rowman & Littlefield.
Shimbori, M. 1960. "A Historical and Social Note on Moral Education in Japan." *Comparative Education Review* 4 (2): 97–101.
Smythe, H. H. 1952. "The Eta: A Marginal Japanese Caste." *American Journal of Sociology* 58 (2): 194–196.
Stepin, V. S. 2006. "Kul'tura i tsennosti v epokhu globalizatsii [Culture and Values in the Era of Globalization]" [in Russian]. *Voprosy filosofii* 5 (3): 3–12.
Stone, R. 2021. "The Middle Path and Pure Experience: A Re-evaluation of the 'Beginning' of Modern Japanese Philosophy." *The Journal of East Asian Philosophy* 1 (1–2): 15–29.
Svensson, M. 2012. "Augustine on Moral Conscience." *The Heythrop Journal* 54 (1): 42–54.
Taira, K. 1969. "Urban Poverty, Ragpickers, and the 'Ants' Villa' in Tokyo." *Economic Development and Cultural Change* 17 (2): 155–177.
Takakusu, J. 1906. "The Social and Ethical Value of the Family System in Japan." *The International Journal of Ethics* 17 (1): 100–106.
The Analects of Confucius. 2007. Trans. from the Chinese by B. Watson. New York: Columbia University Press.
The Chinese Classics: Translated into English with Preliminary Essays and Explanatory Notes by James Legge. 1869. Trans. from the Chinese by J. Legge. London: N. Trübner.
Uccello, I. 2024. "Crossing Cultural Boundaries: The First Translation of Crime and Punishment in Italy." *Translation Studies: Theory and Practice* 4 (1): 74–83.
Wakabayashi, J. 2012. "Japanese Translation Historiography: Origins, Strengths, Weaknesses and Lessons." *Translation Studies* 5 (2): 172–188.

Kravtsov A. D. [Кравцов А. Д.] Morality without a Subject [Мораль без субъекта] : Confucian-Buddhist Foundations of Ethics in the Japanese Translation of Dostoevsky's "Crime and Punishment" [конфуцианско-буддийские основания этики в переводе «Преступления и наказания» Ф. М. Достоевского на японский язык] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 221–241.

Андрей Кравцов
аспирант, Институт философии РАН (Москва); orcid: 0009–0006–1065–6136

Мораль без субъекта

конфуцианско-буддийские основания этики в переводе «Преступления и наказания» Ф. М. Достоевского на японский язык

Получено: 06.09.2025. **Рецензировано:** 26.10.2025. **Принято:** 01.11.2025.

Аннотация: В статье исследуется феномен культурного трансфера литературного произведения как сложного процесса семиотической адаптации, в котором сталкиваются не только языковые структуры, но и глубинные онтологические парадигмы. На материале японского перевода романа Ф. М. Достоевского «Преступление и наказание» периода Мэйдзи анализируются механизмы трансформации христианско-экзистенциальной проблематики оригинала под влиянием буддийско-конфуцианского синтеза. Исследование фокусируется на семантических кластерах («страдание», «совесть», «судьба») и их онтологическом перекодировании: от христианского провиденциализма к буддийскому учению о пустоте, от экзистенциальной рефлексии к конфуцианской этике долга. Методологическая рамка сочетает корпусный анализ с принципами сравнительной философии, вводя понятия «семантической плотности» и «культурного индекса» как количественных маркеров ценностных приоритетов. Автор демонстрирует, как русский экзистенциализм, встречаясь с дзэнской концепцией небытия, порождает гибридные формы: «отчаяние» Раскольникова переосмысляется через отрешенность, христианская «совесть» трансформируется в конфуцианскую врожденную добродетель, а линейная темпоральность романа растворяется в циклической модели непостоянства времени. Философская значимость исследования заключается в раскрытии перевода как творческого акта созидания новой философской реальности, где диалог идет не на уровне заимствований, а через глубинное преображение смыслов. Переводческая практика эпохи Мэйдзи предстает как пространство рождения гибридных онтологий, отражающих сложный процесс модернизации японского общества через синтез традиционных ценностей и западных влияний.

Ключевые слова: культурный трансфер, семантические кластеры, сравнительная философия, корпусная лингвистика, гибридные онтологии.

DOI: 10.17323/2587-8719-2025-4-221-241.

Переводы и публикации

Publications and Translations

Gurov, O. N. 2025. "Beyond Boundaries: A Conversation with Stelarc on His Vision of Human-Machine Integration" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 245–264.

Oleg Gurov*

Beyond Boundaries**

A Conversation with Stelarc on His Vision of Human-Machine Integration

Abstract: For over fifty years, Stelarc has radically explored the fusion of human biology and technology through his provocative artistic practice. This paper draws on a 2024 interview with the artist and examines how his work challenges our fundamental assumptions about human identity and capability. Through analysis of his pioneering projects — from the iconic Third Hand to the Ear on Arm and his internet-enabled performances — we see how Stelarc's art embodies emerging posthumanist and transhumanist philosophies. His artistic work addresses the idea that humanity has moved beyond natural selection into an era where technological engineering drives our development. The research traces his artistic evolution from early body suspensions to recent explorations of distributed embodiment and AI integration, including insights from our 2024 conversation. By merging art, philosophy, and technological innovation, Stelarc creates not just theoretical frameworks but visceral demonstrations of possible human futures. His bold experiments spark crucial dialogue about consciousness, embodiment and identity in our world, which gets to become more and more technological.

Keywords: Stelarc, Cyborgization, Posthumanism, Transhumanism, Body Modification, Performance Art, Human-Machine Interface, Distributed Agency.

DOI: 10.17323/2587-8719-2025-4-245-264.

These days the boundaries between human and machine grow increasingly fluid, and no artist embodies this transformation more literally than Stelarc (Stelios Arcadiou). This research delves into the mind and work of an artist who has spent over half a century challenging our preconceptions about the human body and consciousness. At the heart of this study lies an exclusive conversation with Stelarc, conducted on September 2, 2024 in a remote mode through the Zoom platform. The purpose of the interview

*Oleg Gurov, MBA, PhD in Philosophy; Research Fellow at the Center for Artificial Intelligence at MGIMO University (Moscow, Russia); Associate Professor at the State Academic University for the Humanities (GAUGN) (Moscow, Russia), o.gurov@inno.mgimo.ru, gourov.oleg@gmail.com, ORCID: 0000–0002–8425–1338.

Acknowledgements: The article was prepared at the State Academic University for the Humanities within the framework of the state assignment of the Ministry of Science and Higher Education of the Russian Federation (topic No. FZNF–2023–0004 "Digitalization and Methods of the Modern Information Society: Cognitive, Physical, Political and Legal Aspects").

was to explore Stelarc's current philosophical perspectives on the relationship between humans and technology, as well as his vision for the future of humanity in the context of technological evolution.

Our conversation with Stelarc reveals new perspectives on his philosophy, creative process, and vision for humanity's technological future. Stelarc's provocative performances and bodily modifications serve as radical experiments that challenge our understanding of human potential. By analyzing Stelarc's boundary-pushing work with cyborgization and human augmentation, we gain crucial insights into how technology reshapes human identity. His artistic research serves not just as art, but as a window into possible futures where the distinction between human and machine becomes increasingly meaningless.

Stelarc is a Cypriot-Australian performance artist and body modification pioneer born in 1946. For over five decades, Stelarc has pushed the boundaries of human physiology, technology, and art, becoming a seminal figure in the field of cyborgization and posthuman exploration (Fistrek, 2024). Since his early work in the 1970s, Stelarc has consistently developed innovative approaches to merging human biology with technological systems, creating functioning prototypes of human augmentation (Baudrillard, Glaser, 1994).

Stelarc's journey into cyborgization began with his radical internal body explorations from 1973 to 1976. Using medical endoscopy equipment, he filmed three meters of his internal space, including his stomach, lungs, and colon. This intimate exploration of the body's interior was really innovative, not only as an artistic statement but as a precursor to his later work in extending and augmenting human physiology.

In the early stages of his artistic journey, Stelarc turned to medical technology to reveal the hidden symphony of the human body. Performing as a scientist-artist hybrid, he transformed clinical tools into instruments of revelation. Brain waves danced across screens through EEG recordings, heartbeats thundered through amplified ECG signals, and ultrasound sensors mapped the rushing rivers of blood beneath the skin. Through EMG readings, even the subtle whispers of muscle movement became visible, turning the body inside-out for all to witness.

The 1980s brought what many consider Stelarc's masterpiece — the "Third Hand." This was more than a prosthetic; as it was a glimpse into humanity's cyborg future. Custom-built and attached to his right arm, this robotic appendage could perform feats beyond natural human capability, including a striking 290-degree wrist rotation. The following made it a really revolutionary project: its control system, the hand responded to EMG

signals from Stelarc's abdomen and leg muscles, was effectively rewiring his body's natural control systems. It can be regarded as a working prototype of human augmentation, demonstrating how technology can extend physical capabilities beyond biological limits.

Handswriting. Maki Gallery, Tokyo (1982). Photograph: Keisuke Oki. © Stelarc.

Each performance operated like a live experiment and demonstrated a human-machine synthesis, with Stelarc's body functioning both as experimental subject and as artistic medium. The work pushed into uncharted territory and challenged notions of what it means to be human. It was innovative in exploring human-machine interfaces and the potential for a closer integration of prosthetic devices with the human body. Through his performances, Stelarc demonstrated novel possibilities for distributed agency and remote embodiment, fundamentally challenging traditional notions of bodily autonomy. They also raised important ethical questions about the nature of identity and agency in a technologically-mediated world.

Stelarc works advance practices in remote control, and wearable biofeedback and have fostered interdisciplinary collaborations between art, science, and engineering. They prompt concrete philosophical questions about durability, access, and the social distribution of enhancement, rather than

making broad claims about human evolution. His work continues to provoke discussion among artists, scientists, and philosophers about posthuman possibilities.

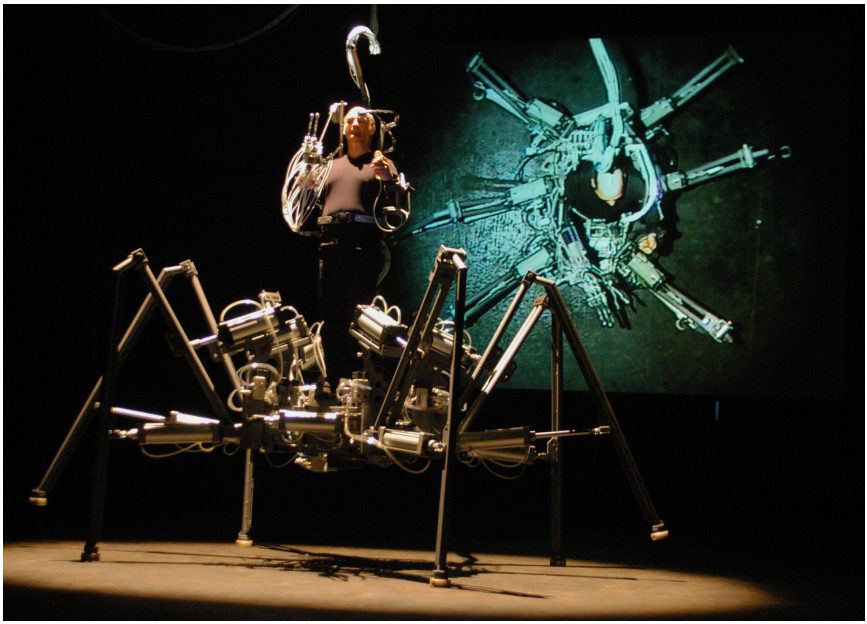
Stelarc's transformative artistic interventions have been generating significant academic discourse across disciplines, from performance studies to bioethics, sparking intense debates about the nature of human embodiment, the increasingly blurred lines between organic and artificial, and the trajectory of human evolution in our rapidly advancing technological landscape. At the heart of Stelarc's oeuvre is a provocative reimagining of the human body as subject to technological augmentation and redesign. This perspective has drawn significant attention from researchers in fields ranging from performance studies to bioethics, cybernetics, and posthuman theory, each bringing their unique lens to interpret and analyze the profound implications of his work.

Central to Stelarc's practice is the concept of cyborgization, a theme increasingly pertinent in the digital age. His famous declaration that "the body is obsolete" serves as a provocative call to embrace technological enhancement as a means of transcending biological limitations (Stelarc, 1991: 593).

The research "Cyborg Art and Bioethics: Stelarc and The Third Ear" by Valeria Radrigán discusses Stelarc's work as exploring the possibilities of extending and modifying the human body through technology (Radrigán, 2013). According to her position Stelarc challenges conventional perceptions of the human body in the technological age. Through his provocative work, he argues that our biological form has become outdated and insufficient for modern technological demands. His creative endeavors involve dramatic corporeal alteration, including his famous suspension performances where he explores the structural capabilities of human skin. As she considered, Stelarc's most controversial piece, the "Third Ear" project, featured a surgically constructed ear on his forearm, designed to function as an Internet-enabled acoustic device. Through boundary-pushing experiments, the artist explores novel anatomical configurations and technical amplification of human capabilities.

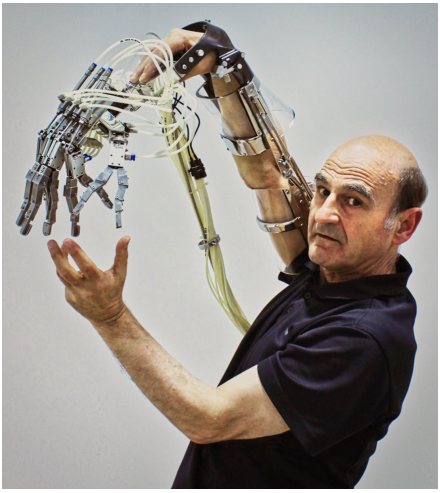
Art theorist Vid Simoniti offers a detailed analysis of Stelarc's artistic philosophy, particularly focusing on his investigation of technological body modification and enhancement (Simoniti, 2019: 177–178). Simoniti examines how Stelarc's performances challenge our understanding of bodily limitations and technological integration. The scholar particularly emphasizes how Stelarc's work, including his suspension pieces and the third ear implant,

serves as a practical demonstration of his theoretical framework regarding human obsolescence and technological necessity.

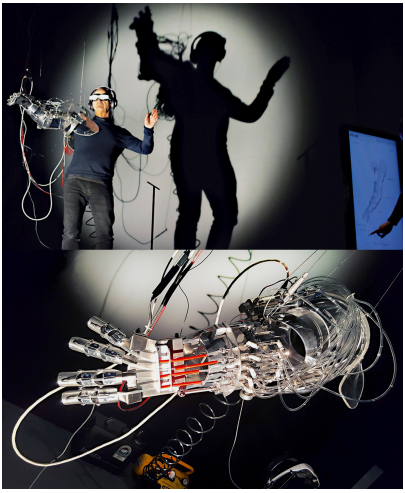


Exoskeleton. Cankarjev Dom, Ljubljana (2003). Photograph: Igor Skafar. © Stelarc.

Catherine Voison approaches Stelarc's work through the lens of bio-art, examining how biotechnology becomes both medium and message in his creative endeavors. She argues that Stelarc's conceptualization of the body as a malleable machine represents a fundamental shift in how we understand human physicality (Voison, 2019: 54). His performances, according to Voison, point toward a post-human future where biological and technological boundaries become increasingly blurred. She introduces the concept of "anthropotechnology" to describe Stelarc's radical physical transformation, suggesting they represent a fundamental reimagining of human nature. Both scholars ultimately view Stelarc's work as a groundbreaking exploration of human-technology synthesis that fundamentally challenges our traditional understanding of embodiment and identity. His performances and body modifications serve as provocative demonstrations of posthumanist ideas about transcending biological limitations through cyborgization.



Extended Arm. Melbourne, Hamburg (2000). Photograph: Dean Winter. © Stelarc.



Re-Wired / Re-Mixed. Radical Ecologies, Perth Institute of Contemporary Art, Perth (2017). Photograph: Steven Aaron Hughes. © Stelarc.

In relatively recent interview with Stelarc taken by Sophia Lawler-Dormer, the artist elaborates in detail on his views on post-humanism and cyborgization (Lawler-Dormer, 2018). In this interview, Stelarc expresses the opinion that the human body in its current form is insufficient and inadequate for modern conditions. He envisions a future in which nanotechnologies will "re-colonize" the body from within, significantly altering its functions and capabilities. Stelarc describes the contemporary human body as a "chimera of flesh, metal and code," emphasizing the already existing fusion of the biological and the technological. The artist also expresses the view that technologies are gradually erasing the traditional distinctions between male and female, which may ultimately lead to the obsolescence of traditional modes of reproduction. He sees the potential of technologies in significantly extending human life, possibly even to infinity, through the replacement of organs with artificial analogues. Stelarc emphasizes that the goal of his work is to explore "alternative anatomical architectures" and broaden the horizons of understanding what constitutes the human body. His experiments are aimed at challenging traditional notions of corporeality and identity, offering radically new possibilities for human existence in the era of technological progress.

Thus, although the sources differ in their emphases and depth of analysis of specific aspects of his work and philosophy, for example, examining Stelarc's works in the context of "bio-art," paying particular attention to the use of biotechnologies to modify the artist's body, or focusing on the influence of the artist's works on perceptions of bodily integrity and individual identity, all sources agree that Stelarc's work represents a radical exploration of the integration of human and technology. The authors unanimously note that his performances and body modifications challenge traditional notions of the human body and identity. Moreover, all sources emphasize that Stelarc views the body as obsolete and inadequate for the contemporary technological environment. There is also a common emphasis on the fact that Stelarc's works demonstrate post-humanist ideas about overcoming biological limitations through cyborgization.

Our conversation with Stelarc allows a comprehensive epistemological reading of the artist's contemporary practice and philosophical stance. The dialogue clarifies Stelarc's sustained inquiry into the limits of corporeality and consciousness against the backdrop of accelerating technological change.

We have attempted to present the Stelarc interview not merely as documentary material that exemplifies these transformations but as a methodological resource of a particular quality: in the artist's self-interpretations the problem vectors highlighted in our theoretical treatment are clearly manifested. Among these themes are hybridity, distributed agency, the politico-ethical questions surrounding bodily modification, and the epistemic function of performance. Because of this, the interview enables the translation of theoretical categories into concrete empirical propositions. This, in turn, makes it possible to identify pressing, practice-relevant issues, including the practical limits of technological bodily modification and the non-obvious but critical risks that such projects entail, which ultimately delineate the boundaries of permissible and feasible intervention (McLuhan, 1964; Stiegler, Beardsworth & Collins, 1998).

From these observations, we draw the study's principal conclusions. First, Stelarc's artistic practices correlate strongly with posthumanist concepts and can be treated as research cases for analyzing technology's relational agency. Second, performance in Stelarc's work functions not simply as demonstration but as a method for producing empirical data about human experience under technological integration and, more broadly, about the social effects of such processes (Simoniti, 2019). Third, Stelarc's projects show that networked interfaces and communication protocols reconfigure perception

and agency, transforming human identity into a multidimensional, extended phenomenon (Hayles, 1999; McLuhan, 1964).

These findings imply that theory divorced from material and practical concerns, especially in the context of complex techno-social bodily processes, remains incomplete. When scaled to political and humanitarian levels, issues such as access entitlements, legal responsibility, cybersecurity, and related matters demand institutional responses (Agamben, Attell, 2004). Declarative ethical codes alone are insufficient because they frequently fail to account for the concrete practices in which these problems arise.

In sum, Stelarc's oeuvre largely affirms central posthumanist intuitions while directing attention to the necessity of applying those theories in practice. His work lays a foundation for a grounded and competent understanding of emergent forms of real human-machine integration, an understanding that combines conceptual rigor with sensitivity to technical, organizational, and perceptual realities (Filas, 2019; Haraway, 1985; Stelarc, 1991; Stiegler, Beardsworth & Collins, 1998).

***

Oleg Gurov: Let's start with the question: What does cyberization mean to you? How would you define this phenomenon both as an artist and as a scientist, since you are much more than just an artist? Is it a reality that is actually taking place in th world, or is it something speculative?

Stelarc: Firstly, the body has always been a prosthetic body — a body that has always been in excess of its biology. From the beginning of human civilization, there was a necessity to construct artifacts and engineer instruments and machines. So, technology has always been coupled with the body. The body has always been prosthetic, as philosophers, including Stiegler have discussed. I lived in Japan from 1970 to 1989. I was quite naive when I went there soon after art school. I hadn't read about or was familiar with the concept of the cyborg. But in Japan, three things really impressed me about the culture: Japanese Sumo wrestling, Butoh dance, and Japanese robotics. These exposed the problematics of what a body is and how it operates — the materiality of the body and how we can be embodied in different ways, whether through a biological body, an athlete, a dancer, or robotic humanoid robots. This exposed me to extreme embodiments and materiality. It also allowed access to technology I wouldn't have had in Australia, like the latest medical, laser and robot technologies. In 1976, for example, I made three films of the inside of my body using endoscopic technology. This would not have been possible had I remained in Australia.

The original meaning of "cyborg" goes back to before the 1970s, maybe to the 1960s. It combined the words "cybernetics" and "organism" — a cybernetic organism. But I think in terms of popular culture, it didn't really come into common use until the 1980s. Then it was understood as a kind of hybrid cybernetic system, a biological body with mechanical parts or machine attachments. I like to counterpoint notions of cyborgs with notions of zombies. A Zombie is a body that performs involuntarily, which does not have a mind of its own. A Cyborg is a human-machine system that becomes increasingly automated. There has always been a fear of the involuntary and an anxiety of the automated. Of the Zombie and the Cyborg. But we fear what we have always been and what we have already become.

O. G.: That's great. I completely agree it really came to mass culture in the '80s. How do you think this concept has changed from the '80s to nowadays? Are there similarities or differences?

S.: Well, much of Donna Harraway's original "Cyborg Manifesto" is still relevant in its challenge to go beyond traditional feminism, the fluidity of identity, an emphasis of an ontology of hybridity and the blurring of boundaries between the human, the animal, and the machine. Since the mid-1980s, cyborg constructs have undergone further nuanced change from a purely mechanical, instrumental kind of embodiment to now even incorporating genetic interventions and surgical modifications. So, the idea of the cyborg is not simplistically machinic, but rather a construct with much broader range of interventions, augmentations and enhancement that can generate what I call a contemporary chimera of metal and code — of biology, technology, and virtuality. With the introduction of AI, you have more complex cybernetic feedback loops occurring that generate a more sophisticated and subtle hybrid machine system. Having said that, the word "Anthropocene," like the word "cyborg" is not so commonly used now. It reached its peak impact maybe a decade or two ago. In terms of critical theory, the word cyborg now has become somewhat banal in its use and interpretation. But it remains a convenient concept to apply to this kind of techno-organic system.

O. G.: In my research, I write about Marshall McLuhan's idea of the extended sensorium, which originated in the '60s or '70s. It implies that humans are constantly evolving, which you also mentioned at the beginning of our conversation. Does your own sensory experience change when you do your technological experiments with your body in your art?

S.: Firstly, I think Marshall McLuhan hasn't received enough credit as a philosopher. I read McLuhan when I was in art school. His work goes back to Whitehead and has influenced Baudrillard and Virilio, who focused on

particular aspects of McLuhan's ideas and developed them further. What interested me about McLuhan is the idea that technology is an extension of the body. But he says more than this — that we externalize and extend our nervous system, and the internet can be seen as a kind of external nervous system of the body. Also, his very seductive statement that technology can be seen as the external organs of the body. We have evolved as biological bodies with soft internal organs, but now, inhabiting a technological terrain of fast and powerful machines, we need to develop additional external organs to better interface with our new technologies. For me, I'm neither utopian nor dystopian about technology. I think technology is the trajectory of human civilization and it is certainly not an alien other. Technology is what has always been coupled to the body and has been primarily responsible for our humanity and our civilization. It's also important to note that evolution is probably the wrong word to use for what's happening now. There's no evidence to indicate that evolution has ceased with the human body, but because evolution is such a slow process, taking millions of years, what's happening now is different. We've gone from the process of evolution to a process of engineering, applying human design ideas and methods, and generating much more accelerated change. Contemporary technological development creates complex systems of mutual influence between biological and synthetic elements, catalyzing rapid transformations in human capability. So, it can be argued that the process now is not a Darwinian evolution but rather one of Lamarkian change.

O. G.: Let's talk about how humans are changing within technological means. What do you think about the change of consciousness and human identity? Can we see some preservation of a coherent self, or is there something else happening?

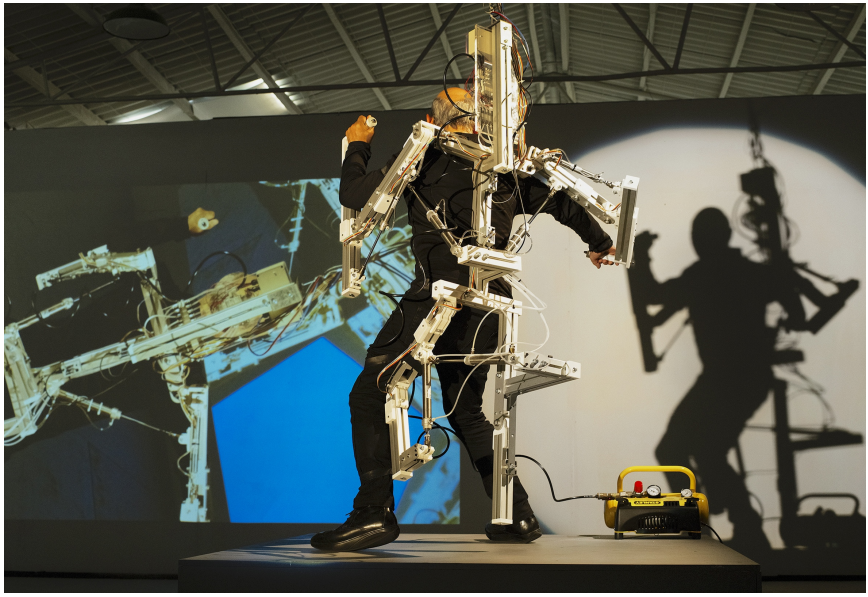
S.: Changes, even though they're accelerating, are happening incrementally and in a multiplicity of different ways. It's not one direction, but a multiplicity of directions and unexpected possibilities. Some will be beneficial, some will not, some will succeed, and others will not. We know that technology alters human behavior. It radically alters our personal habits and social behaviors. Take the example of wireless technologies like mobile phones. This has allowed us a certain mobility that comes with connectivity, allowing us to project our physical presence and instantly connect with people in other places in ways we couldn't before. We're extruding our sense of self beyond our skin and beyond the proximal location we inhabit. In more significant ways, because of this dilemma of what a body is and how it operates, and ambivalence about gender, we've developed surgical

techniques that can even modify our gender. With CRISPR technology, we can now intervene genetically more effectively and accurately altering our DNA. These technologies change us both psychologically and physically. The human construct is somewhat unstable and historically positioned. What it means to be human today might not have been considered human even a few hundred years ago. What it means to be human now will be very different in a thousand or two thousand years. We might look the same, we might have a similar form with similar functions, but all technology in the future might be invisible because it's micro-scaled and nano-scaled, inside the human body. We will be able to recolonize the human body, augmenting our bacterial and viral population. The body, once contained by technology, now incorporates it as a component.

O. G.: You've already answered a question regarding the body and technology. I also wanted to include one more factor such as ESG, which is very relevant in today's world. What about the relationship between the body, technology, and the environment? Can we speak of the emergence of new forms of environmental consciousness? Is it real science or fiction? If it's real, in what forms? How do you reflect this in your work?

S.: I've always been skeptical about subjective reporting and speculating. What I have to say is determined by my projects and performances, and of course what I've read, which is often prompted by what I do. For example, with the Rewired Remixed performance, for five days, six hours every day, I could only see with the eyes of someone in London, I could only hear with the ears of someone in New York, but anyone anywhere could access my right arm and remotely choreograph its movements. It was a kind of distribution of agency to remote people in other places, but also sharing of visual and acoustical senses. The body was virtually in two places and physically in another place at the same time. The performance was done with a kind of posture of indifference — not having any expectations, allowing the performance to unfold in its own time with its own rhythm, and just experiencing the possibilities. I didn't know what I was going to hear or see at any moment, or when someone was going to throw me off balance by moving my right arm. This was a performance where you tried to experience what it means to be a body that performs beyond its skin and beyond its subjective agency. These are aesthetic experiences rather than illustrations of a particular idea. For me, the definition of art is the slippage that happens between what the artist intends and what the actual outcome is. This slippage between intention and outcome is the realm of art, which can incorporate the accidental and the unexpected. Generally, people now

are operating in radically different ways than they were even 50 years ago. We take the internet for granted, but it's a very potent external nervous system. It connects bodies, (nodes, these bodily nodes,) brains, arms, and hands, and it allows collaboration between these nodes of interactivity, which are human bodies connected to terminals globally. Teilhard de Chardin's concept of the noosphere is relevant here. As well as our biosphere, we now have a Chardinian noosphere — a mental sphere. Marshall McLuhan's notion of the "global village" also comes from that. And this is really what the Internet has become, this kind of additional operational, cognitive and collaboratively interactive layer between bodies, between humans. But we need to remember that the digital is not only about the virtual, is not only about information archives but rather it allows for alternative embodiments. In an age of information overload what becomes important is not freedom of information but rather freedom of form, freedom to modify your body.



StickMan. Chrissie Parrot Arts, Perth (2017). Photograph: Toni Wilkinson. © Stelarc.

O. G.: Stelarc, what you are talking about sounds very provocative, even now when I think no one can be surprised by anything. You are a world-famous artist who does such radical artistic performances, which involve

things that were previously prohibited or just not really common, connected with extension of the body in public. How would you describe your art or performances? And how does it help us, the people who are watching and admiring you, to overcome the fear of human-machine fusion? Or, on the contrary, do you just want to attract attention to this and make people feel responsible?

S.: To give a specific example of the physical changes and the changes in our psychology and social and cultural acceptances now, I was asked to give a keynote at a body hacking conference in Austin, Texas. There was a woman who attended who had an artificial leg. This artificial leg was 3D printed with an intricate and aesthetic design. She didn't hide it and was very proud of the fact that she had an artificial leg. She didn't try to make it cosmetic in appearance. Another person, a male, had an artificial arm made of carbon fiber and aluminum, but he didn't cover it. He was quite proud of it and wanted to show off his cyborgian arm. He could even remove the hand and put on different attachments. These examples show that increasingly now, socially, we accept bodies that are patched up with prosthetic replacements. Not only are we comfortable socially and culturally accepting them, but these prosthetically augmented individuals are not self-conscious. They're quite proud of being part biology, part technology. In terms of what these projects and performances do, firstly, one must understand that art is not about illustrating or demonstrating some kind of ideology. Art is not ordinarily about any kind of propaganda, whether political or ideological. What these projects and performances do is expose the problematics of what a body is and how it operates, the hybridization of the body with technology. Interestingly now, there are two parallel happenings. On the one hand, we have robots becoming more and more human-like, in their machine musculature and dexterity. On the other hand, we have bodies that are becoming more and more machinic and automated in their behavior. At a certain point in time, perhaps these two trajectories will converge where it's going to be meaningless to distinguish between the two. If a robot speaks like me, looks like me, is socially adept and responds to unpredictable situations like me, who am I to deny its intelligence or even its ethical rights in the world? The philosopher Daniel Dennett indicates that you can have competence without comprehension. We do not need to attribute consciousness and feeling to robots if they can effectively emulate human behavior in the world.

O. G.: Stelarc, as usual, you predicted my next question. Thank you very much. I wanted to ask your opinion regarding personal boundaries.

In your experiments as an artist, do you have any limitations that must be present in your art?

S.: Well, the process of art is a more intuitive method that is more about affect than being informative. That generates an ambivalence, an uncertainty and sometimes even an anxiety. That is not about providing solutions but keeps asking questions. As a performance artist, I don't sit down and think, "What radical action will I do next?" Rather, each performance generates other alternative possibilities. As a performance artist, what's meaningful is not simply to speculate, but to actualize an idea, to perform the idea, to experience it and hopefully have something meaningful to articulate afterwards. For example, with the Ear on My Arm project, people ask, "Why not an eye?" But to engineer an extra eye, at this point in time, with our technologies and medical expertise, it's just not possible. But to construct an ear that is an external structure of an ear was plausible. It was pushing the boundaries, but it was plausible. It took 10 years to find three surgeons and to get funding to do the first surgery. I'm only interested in actualizing ideas that are plausible, that are possible. I'm not just interested in speculating. I think it's easy to have an idea. What's difficult is to actualize the idea and to physically experience it. But the artist has to take the consequences for those ideas. If you want to suspend your body, you have to stick 18 hooks into your skin. Or if you want to insert a sculpture or a machine inside your stomach, there are medical problems and risks to take. I think we have to think of boundaries in a different way now. In this present age of Circulating Flesh, Fractal Flesh, and Phantom Flesh, we shouldn't think of being bounded so much as occupying transitory states. We're in liminal states of transition. A boundary is no longer applicable if we're performing beyond the boundaries of our skin, if we're performing beyond the boundaries of the local space that we inhabit, if we perform beyond the boundaries of our biosphere. We're no longer defined by our boundaries. We're now in an age of liminality.

O. G.: Thank you very much. The next question is extremely theoretical, but with your scholarly background, I think it will be easy for you. I wanted to ask you about the post-humanistic and transhumanistic approaches. From my point of view, the post-humanistic approach means that humanity is going to overcome humanity and become something else. The transhumanistic approach means that we, as humans, are developing within technological means. In my work, I try to research your art, and at one point I think you are a transhuman artist, and at another point, I think you promote other

ideas. What do you understand by this, and in your art, what do you want to express? In what state is humanity moving?

Reclining StickMan. Monster Theatres, Biennial of Australian Art, AGSA. Adelaide (2020). Photograph: Saul Steed. © Stelarc.

S.: Oh, there is no desire to promote particular ideas nor to be categorized as Transhumanist or Posthumanist. Categories are convenient but can also prove simplistic. What might be meaningful is whether we prioritize the human species as such or we perpetuate more intelligent life-forms. If our post-evolutionary direction is to perpetuate intelligence then this vulnerable body in this form and with these functions might not be the best way of achieving it. Is intelligence better perpetuated in some kind of redesigned, reimagined, differently embodied intelligent agent? For example, in some hybrid human machine, robot or chimera or whatever. If our goal as an intelligent species is to perpetuate intelligence, to guarantee intelligence is not only disseminated but survives, then perhaps remaining on this planet is a bad survival strategy. Perhaps it's necessary to go beyond this body and go beyond this planet. Taking a more immense timescale, and if we realize that all living things are destined to vanish forever, then perhaps in realizing this, we should plot an elegant exit. What's important is not necessarily

the survival of the human species, but if evolution does have a higher purpose, it's to generate and disseminate higher forms of intelligence.

O. G.: That's nice to hear that you are also optimistic about it. What do you think about artificial intelligence in this meaning? What do you think about the prospects for integration of human intelligence within artificial intelligence systems? You already mentioned this, but what do you think are the risks and opportunities within such a collaboration?

S.: The kind of research now is developing an artificial general intelligence, so that this AI is not only capable of performing particular tasks, but can interact with the world, learn from its interactions, and perhaps develop in some interesting way. For me, what would be meaningful about artificial intelligence is if it ever becomes an alien intelligence. I don't mean that in some kind of dystopian way, but it would have certain consequences that we would have to accept and become complicit with. Do we privilege human intelligence or are willing to incubate a machine intelligence that performs with greater cognitive capacity and with greater processing speed and with instantaneous access of vast online archives of information can be much better pattern recognition and analysis. With our 1400cc brains, with our metabolic speed, with our malfunction memories and unreliable retrieval we would not be able to match the performance, nor even comprehend information generated by our AI computational systems. This might be worrisome and sounding dystopian for people who are obsessed with perpetuating the human species, but again, if the human species is a kind of step towards a more intelligent agent, then I see this as a positive direction, a positive future. Of course, we're speculating, for if there is a future, there are scientific theories, like the block universe, where past, present, and future are existing simultaneously. William Gibson's statement that "the future is already here, it's just not equally distributed" is relevant here. Of course, he's also alluding to issues of access, priority, and privileging, but in terms of the construct of time-space, it doesn't have to be this linear progression of the past to the present and onto the future. We can characterize the past as memory, the present as understanding, and the future as imagining. So, we can characterize past, present and future as a purely human construct, as a purely bodily construct. In fact, the philosopher Kant points out that time and space don't exist objectively, they are not real properties of the world. Rather they are categories of human understanding. Time and space are the means by which the body experiences reality. As a forward-facing body with two eyes and two ears to navigate, and two hands and limbs to manipulate with, we construct the self

as an intentional agent. And being forward facing the body subjectively is future oriented. The self-experiences the world as what it has forgotten, what it now comprehends and what it might imagine.

O. G.: If space and time are changing within all these processes we're talking about, what do you think new art forms or new kinds of art can emerge in this situation? I mean, in the next decades, because it's obvious that humans will be more and more immersed in virtuality and with artificial intelligence. How will it affect the arts?

S.: Again, speculating, but if we look at the present, I'm performing with prosthetics, I'm performing with exoskeletons, I'm performing with robots, I'm performing remotely and interactively online, I'm performing with actual-virtual interfaces. So, I think new instruments, new machines are always seductive for artists because they generate new conceptual and aesthetic possibilities that can be problematized and explored. There is this feedback loop between the zeitgeist, the technologies of the time, and the generation of new artists who come up with unique ideas. This will result in unexpected art forms and modes of expression. I don't necessarily think that the digital age will only be a virtual age. The digital has also resulted in humanoid robots and surgical sex change operations. We have to be open to different possibilities that will be both conceptually surprising and aesthetically novel. Art is not interesting if it's not surprising in some way. And it's the same with the future — if the future can be predicted, then by definition, it's not a future at all. The future is not a future if it is not of the unexpected. So, we may see art forms emerging that blend the physical and virtual in new ways, that incorporate AI and robotics, that play with notions of embodiment and consciousness. But the specific forms are impossible to predict — that's what makes the future of art exciting. Art is not what happens of necessity but rather what is contingent and contestable.

O. G.: Thank you for those fascinating insights into how technology might shape future art forms. As we wrap up, is there anything else you'd like to add about your vision for the future of art and the human body?

S.: I think the key is to remain open to possibilities we can't yet imagine. As artists, we need to keep taking conceptual and physical risks, asking difficult questions, and exploring alternative modes of embodiment and expression. It is time to question whether a bipedal, breathing body with binocular vision and a 1400cc brain is an adequate biological form. The body is neither a very efficient nor very durable structure. It malfunctions often and fatigues quickly; its performance is determined by its age. It is susceptible to disease and is doomed to a certain and early death. Its survival

parameters are very slim — it can survive only weeks without food, days without water and minutes without oxygen. The body's lack of modular design and its overreactive immunological system make it difficult to replace malfunctioning or failed organs. In the early 1970s I wrote that the body artist of the future will be a genetic sculptor. We need to reconsider how we conceive of the body, consciousness, and reality itself. The dead, the brain dead, the yet to be born, the cryogenically preserved, the prosthetically augmented, the plasticated, synthetic life and artificial life all now share a material and proximal existence. Art has a vital role to play in helping us navigate and make sense of these unexpected juxtapositions. So, we must continue to experiment, to provoke, and to imagine new ways of being bodies in a world of rapidly evolving technology that radically shapes our humanity. Being human is perhaps not remaining human at all.

O. G.: That wraps up our talk on a meaningful note. Stelarc, I can't thank you enough, this has been really an eye-opening conversation. Your work keeps sparking new ideas and inspiring people, whether they're seasoned artists or just discovering your work.

S.: Thank you, I enjoyed it too. I am excited to see what the up-and-coming artists will do with all this. I wish they will come up with ways to explore these themes that we haven't even imagined yet.

REFERENCES

Agamben, G. 2004. *The Open [L'Aperto]: Man and Animal [L'uomo e l'animale].* Trans. from the Italian by K. Attell. Stanford (CA): Stanford University Press.

Baudrillard, J. 1994. *Simulacra and Simulation [Simulacres et simulation].* Trans. from the French by S. F. Glaser. Ann Arbor (MI): University of Michigan Press / Semiotext(e).

Filas, M. 2019. "My Dinner with Stelarc: A Review of Techno-flesh Hybridity in Art." *The Information Society: An International Journal* 35 (4): 188–199.

Fistrek, L. 2024. "Stelarc and the Post Human Body Performance in Contemporary Art." Academia.edu. Accessed Dec. 2, 2024. https://www.academia.edu/34546795/STELARC_AND_THE_POST_HUMAN_BODY_PERFORMANCE_IN_CONTEMPORARY_ART.

Haraway, D. J. 1985. "A Manifesto for Cyborgs: Science, Technology, and Socialist Feminism in the 1980s." *Socialist Review*, no. 80, 65–108.

Hayles, N. K. 1999. *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics.* Chicago: University of Chicago Press.

Lawler-Dormer, S. 2018. "Redefining The Human Body As 'Meat, Metal and Code': An Interview with Stelarc." Sleek Magazine. Accessed Dec. 2, 2024. https://www.sleek-mag.com/article/stelarc-interview-posthumanism/.

McLuhan, M. 1964. *Understanding Media: The Extensions of Man.* New York: McGraw-Hill.

Radrigán, V. 2013. "Cyborg art y bioética: Stelarc y The third ear" [in Spanish]. *Aisthesis* 54.

Simoniti, V. 2019. "The Living Image in Bio-Art and in Philosophy." *Oxford Art Journal* 42 (2): 177–196.

Stelarc. 1991. "Prosthetics, Robotics and Remote Existence: Postevolutionary Strategies." *Leonardo* 24 (5): 591–595.

Stiegler, B. 1998. *The Fault of Epimetheus [La faute d'Epiméthée].* Vol. 1 of *Technics and Time [La technique et le temps]*, trans. from the French by R. Beardsworth and G. Collins. Stanford: Stanford University Press.

Voison, C. 2019. "L'art biotechnologique: une anticipation d'un au-delà de l'humain?" [In French]. *Journal International de Bioéthique et d'Éthique des Sciences* 30 (4): 51–68.

Gurov O. N. [Гуров О. Н.] Beyond Boundaries [Преодоление границ] : A Conversation with Stelarc on His Vision of Human-Machine Integration [беседа со Стеларком о его видении интеграции человека и машины] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 245–264.

Олег Гуров
MBA, к. филос. н., научный сотрудник, Центр искусственного интеллекта МГИМО МИД России (Москва); доцент, ГАУГН (Москва); orcid: 0000-0002-8425-1338

Преодоление границ

беседа со Стеларком о его видении интеграции человека и машины

Аннотация: На протяжении половины столетия художник Стеларк проводит радикальные художественные эксперименты, разнообразными способами совмещая аспекты человеческой биологии с технологическими системами. Данная работа, в основе которой лежат материалы эксклюзивного интервью с художником, взятого в сентябре 2024 года, а также результаты анализа его творчества, представляет собой попытку систематизировать художественные практики Стеларка с точки зрения динамики определения человеческой идентичности и возможностей человека. Путем детального исследования его различных проектов и перформансов показывается, как работы Стеларка воплощают развитие постгуманистической и трансгуманистической философии. Радикальный взгляд художника направлен на то, что человечество уже вступило в постэволюционную фазу, в которой развитие человека обусловлено в большей степени технологическим проектированием, нежели естественным отбором. В рамках исследования сделан обзор художественного пути Стеларка, его творчество представлено в хронологическом порядке: от ранних перформансов с подвешиванием тела до современных роботизированных инсталляций, что позволило наглядно показать последовательность исследовательских проектов художника в части изучения разных аспектов интеграции человека и маши-

ны. При этом отдельное внимание уделяется недавним работам Стеларка, посвященным проблематике распределенной телесности и интеграции искусственного интеллекта в общественную жизнь и природу человека. В статье показано, что творчество Стеларка объединяет искусство, философию и технологические инновации беспрецедентным образом, в результате чего новые концепции сопровождаются осязаемой и нередко провокационной демонстрацией возможных перспектив «человеческого». Делается вывод, что посредством смелых перформансов и физических модификаций Стеларк вносит вклад в междисциплинарную дискуссию о сознании, телесности и идентичности человека в мире, становящемся все более технологизированным.

Ключевые слова: Стеларк, киборгизация, постгуманизм, трансгуманизм, модификация тела, перформанс, человеко-машинный интерфейс, распределенная агентность.

DOI: 10.17323/2587-8719-2025-4-245-264.

Философская критика

Book Reviews

Da Mata, J. V. T. 2025. "The Development of Vasiliev's Ideas and Paraconsistent Logic in Russia and Outside: A Review of the Second Russian Edition of Vasiliev's 'Imaginary Logic'" [in English]. *Filosofiya. Zhurnal Vysshey shkoly ekonomiki [Philosophy. Journal of the Higher School of Economics]* 9 (4), 267–274.

José Verissimo Teixeira da Mata*

The Development of Vasiliev's Ideas and Paraconsistent Logic in Russia and Outside**

A Review of the Second Russian Edition of Vasiliev's «Imaginary Logic»

Vasiliev, N. A. 2025. *Voobrazhayemaya logika. Izbrannyye trudy [Imaginary Logic. Selected Works]* [in Russian]. Ed. by V. A. Bazhanov. Moskva [Moscow]: Kanon+

DOI: 10.17323/2587-8719-2025-4-267-274.

The second edition of *Imaginary Logic. Selected Works* of N. A. Vasiliev was published in 2025 by Kanon+ Publishers in Moscow under the direction of Professor Valentin A. Bazhanov. The first edition was prepared by Vladimir A. Smirnov in 1989 (Vasil'ev, 1989).

This revised and expanded edition includes a new preface by Professor Bazhanov, as well as the preface from the first edition by Professor Smirnov. Bazhanov reminds us who Vasiliev really was and what he actually did, as well as why his work as a historian, psychologist, literary scholar, symbolist poet, and translator remains of interest.[1] In his preface, Bazhanov states that Vasiliev anticipated some important keys to non-classical logics. He also emphasizes Vasiliev's central significance for a group of thinkers in Russia who contributed significantly to the logical-gnoseological studies that constitute a prominent branch of that country's philosophy.

Vasiliev was a man ahead of his time, and his works only gradually became known as the horizon of logic broadened. Internationally, the Fifth

*José Verissimo Teixeira da Mata, Professor; Brazilian Logic Society (Brasilia, Brazil), joseverissimo@terra.com.br.

[1] Vasiliev completed medical school at the University of Kazan and later graduated from the faculty of history, as Smirnov tells us in his essay (Smirnov, 1989a). Bazhanov, for his part, focuses on Vasiliev's intellectual legacy.

International Philosophical Congress, held in Naples in 1924, allowed for a brief communication in English of the conception of a logic without the principle of non-contradiction and without the principle of the excluded middle. This Naples communication would give some international resonance to Vasiliev's name, because in 1936 his theses were listed in Alonzo Church's celebrated bibliography of symbolic logic (Church, 1936). It was only some decades later, in 1965, that Smirnov — the first Soviet logician to analyze Vasiliev's logical thought in depth — saw his essay on the interpretation of Vasiliev's logic reviewed by David Dinsmore Comey in the *Journal of Symbolic Logic* (Comey, 1965).

Bazhanov reviews the various ways in which attempts have been made to situate Vasiliev within the history of non-classical logics. Kline positioned Vasiliev as a precursor of polyvalent logics (Kline, 1965). The great Soviet algebraist A. I. Maltsev (Mal'tsev, 1976: 474–475), for his part, had already noted that some constructions based on Vasiliev's theses can be understood as modal. For some, Vasiliev, along with Orlov, should be considered one of the inspirers of relevance logics (Bazhanov, 2007: 244–277). Bazhanov also recalls that Smirnov classified Vasiliev's logic as multidimensional, insofar as this logic could produce different and new types of judgments. Vasiliev distinguishes three different structures of judgments: affirmation, negation, and indifferent judgment (here one could even consider a set with different extensions of negation that are associated with different types of negative or indifferent judgments). These distinct structures are considered as different kinds of dimensions. In his preface, Smirnov also points out the importance that Vasiliev's ideas have for dual-intuitionist logic, even if indirectly. If the principle of the excluded middle is not demonstrable in this type of intuitionist logic, then in an anti-intuitionist logic the conjunction $A\&\neg A$ does not constitute a contradiction.

In their prefaces Bazhanov and Smirnov both highlight the fact that the negation of the principle of non-contradiction is central to Vasiliev's thought, and the former defines this idea as the quintessence of the Kazan logician's contribution. Both also assert that, if one considers non-classical logics, Vasiliev's most important connection is to paraconsistent logic, of which he was a precursor. It is true that Smirnov, in his English "The Logical Ideas of N. A. Vasiliev and Modern Logic" (Smirnov, 1989b) and Russian "Multidimesional Logic,"[2] argued that linking Vasiliev to polyvalent, intu-

[2]Smirnov, 1993: 260. Professor D'Ottaviano, in her preface to the posthumous book by Professor Ayda Arruda (Arruda, 1990), notes that Priest and Routley, in a chapter published

itionistic, or paraconsistent logics would not be entirely accurate, despite the fact that he was one of the first to proclaim and construct a non-Aristotelian logic. However, in his preface to the first edition of Vasiliev's works, Smirnov somewhat adjusts his understanding of Vasiliev's relationship to contemporary logical thought and notes that the interest in his logical texts relates primarily to paraconsistent logics.

In this regard, the opening of Smirnov's preface is very telling, and in the very first paragraph he says the following:

> One of the points of advancement in contemporary logical science consists in the study of logical systems in which statements that contradict one another can be formulated and correctly used. There are a number of approaches that motivate the introduction of new systems of this kind, but at the center of attention there is the possibility of expressing contradiction within them. These are above all the C_n systems of the Brazilian logician da Costa, who studies the extension of classical logic with complementary negation, so that statements of the type $A\&\neg A$ are not lost and are not always considered false (Vasiliev, 2025: 12).

Further on, Smirnov addresses the importance of paraconsistent logic for contemporary times:

> The interest in paraconsistent logics is now enormous. They have theoretical significance — for the analysis of contradictory statements, logical and semantic antinomies, and the localization of contradictions — as well as having practical significance insofar as, in principle, different and even contradictory information can enter information retrieval systems. Ultimately, this contradictory information should not destroy the system but should remain localized (ibid.: 14).

In his preface, Smirnov also notes that Bazhanov discovered two important reports by Vasiliev: one on his scientific activity, and another on his work trip abroad.

Bazhanov, to whom we owe this excellent second edition, and who, as noted earlier, perceived Vasiliev's rejection of the principle of (non-)contradiction as the quintessence of his logical thought, says in his preface:

> It seems that Ayda Arruda was the first to draw attention to Vasiliev's rejection of the principle of non-contradiction and to his logic free from this law; she was a disciple of Newton da Costa, who had asked her to study the works of the Russian scientist. Quite accurately, Ayda Arruda and Newton da Costa

in 1989 in the book *Paraconsistent Logic, Essays on the Inconsistent*, suggest that Vasiliev "could also be considered, along with MacColl and Lewis, as one of the founders of intensional logics" (D'Ottaviano, 1990: xv).

considered it appropriate to call Vasiliev a precursor of paraconsistent logic (Bazhanov, 2025: 7).

In the text of his introduction, Bazhanov also cites works that he considers most important in the discussion of issues related to imaginary logic. In this regard, he highlights important names in Soviet mathematics, such as N. N. Lusin and A. I. Maltsev, and also refers to the relevance of the pioneering works of Newton Carneiro Affonso da Costa (Costa, 1993) and Ayda Arruda (Arruda, 1990), which he cites in Portuguese. He also mentions the recently edited book by V. I. Markin and D. V. Zaitsev, entitled *The Logical Legacy of Nikolai Vasiliev and Modern Logic,*[3] which includes articles by Professor Ítala Maria Loffredo D'Ottaviano and Professor Evandro Luís Gomes, Professor Juliana Bueno-Soler and Professor Walter Carnielli, and Professor Otávio Bueno. He also refers to a work by the Brazilian logician João Marcos, published by Unicamp in 2005 (Marcos, 2005). In the appendix of the edition, there appears the article by Ayda Arruda, originally published in English and entitled *On the Imaginary Logic of Vasil'ev* (Arruda, 1979), which has been translated into Russian by V. V. Anosova, a researcher whose thesis, published in 1984, deals with the relationship between imaginary logic and paraconsistent logical systems (Anosova, 1984). There are chapters of Russian and non-Russian logicians, such as V. L. Vasyukov, V. A. Bazhanov, G. V. Sorina, G. Priest, etc.

While the article by Ayda Arruda cited above appears in Russian in Bazhanov's edition, it is worth remembering that the Brazilian logician, when working on her text, envisioned a book containing her own introduction and the translation from Russian to Portuguese of Vasiliev's three main articles. This book, which includes her essay "N. A. Vasiliev and Paraconsistent Logic" and the translation of Vasiliev's articles,[4] was published posthumously by Unicamp in 1990 as the seventh volume of the CLE Collection, edited and prefaced by Professor Ítala Maria Loffredo D'Ottaviano. In her essay, Ayda Arruda states, regarding the primordial relationship of imaginary logic with paraconsistent systems:

[3]Markin & Zaitsev, eds., 2017. This edition, with chapters by many Russian and non-Russian authors, also contains works by the following ones: G. V. Sorina, E. D. Smirnova, V. A. Bazhanov, W. Stelzner, G. Priest, V. L. Vasyukov, I. B. Mikirtumov, J. Y. Beziau, J. V. T. da Mata, V. M. Popov and V. O. Shanguin, V. I. Markin and D. V. Zaitsev.

[4](1) "Sobre os Juízos Particulares, o triângulo das oposições e o princípio do quarto excluído;" (2) "A lógica Imaginária;" (3) "A Lógica e a Metalógica."

> ...we believe that any formalization of Vasiliev's imaginary logic leads to a paraconsistent logic. Whether this will also be a polyvalent logic is a matter of detail or interpretation (Arruda, 1990: 13).

CONCLUSION

This second Russian edition of Vasiliev's works comes at an opportune time and shows that Vasiliev's logic continues to inspire researchers in Brazil, Russia, and the rest of the world. It includes Vasiliev's three most substantial works: (1) "On Particular Judgments, the Triangle of Oppositions, and the Principle of the Excluded Middle;"[5] (2) "Imaginary Logic" (under this title are included the famous essay, as well as the theses from the Naples Congress and the text of a lecture on the subject at Kazan University); and (3) "Logic and Metalogic."

The edition clearly highlights the fruitful relationship that can be established between Vasiliev's ideas and paraconsistent systems.

The book also contains ancillary texts for understanding Vasiliev, such as a report of his scientific activities (an important manuscript discovered by Bazhanov) and a significant philosophical work on ethics. The latter work even has a logical bias in its title, "The Logical and Historical Methods in Ethics." It deals with the differences between the ethical conceptions of Vladimir Soloviev and Leo Tolstoy, to which the Kazan thinker applies notions of system, identifying the fundamental propositions (principles) on which these thinkers based their ideas. There are also reviews by Vasiliev, including reviews of Francis Paulhan's *La Logique de la contradiction* (Paulhan, 1911), of Henri Poincaré's *Dernières pensées* (Poincaré, 2013), and even a review of *Die Prinzipien der Logik* (Von Windelband & Ruge, eds., 1912), which brings together articles on the principles of logic by important authors of the time.

Let us not forget that Bazhanov has added some of Vasiliev's symbolist poems to the book, and the final result of this edition is very good. Vasiliev is no longer just the prolific thinker from Kazan, but a man with multiple

[5]One of the issues discussed in this article, as the title indicates, is the meaning of particular judgments. Vasiliev sees two possibilities: (1) some, but not all, are Y; (2) some are Y. The latter would be very close to universal judgments. Smirnov addresses these issues in his "The Logical Ideas of N. A. Vasiliev and Modern Logic," cited above. In his symbolic reconstruction of Vasiliev's ideas (syllogistic) on judgments, Smirnov introduces the operator T (only some S are P), $T\mathbf{SP}$. In this way, he pays homage to the Russian language, where the word *tol'ko* means "only."

interests, thoughts, and feelings, who by means of only a few texts definitively entered the history of logic. An extraordinary example of this fact is the aforementioned communication in English ("Imaginary Logic") presented by Vasiliev to the Naples Congress in 1925, and which Alonso Church, one of the pioneers of computational theory and Turing's professor at Princeton (1936–1938), with his intuition for the new, immediately picked up and included in his famous 1936 bibliography of symbolic logic.

Finally, it can be concluded from the significant presence of works by Brazilian researchers in the editors' prefaces that Vasiliev's ideas and the study of paraconsistency in Brazil and Russia are in complete synergy.

REFERENCES

Anosova, V. V. 1984. "Logicheskiye idei N. A. Vasil'yeva i paraneprotivorechivyye sistemy logiki [Logical Ideas of N. A. Vasiliev and Paraconsistent Systems of Logic]" [in Russian]. PhD diss., MGU im. Lomonosova [Lomonosov Moscow State University].

Arruda, A. I. 1979. "On the Imaginary Logic of Vasil'ev." In *Proceedings of Fourth Latin American Symposium on Mathematical Logic*, 1–41. Amsterdam: North Holland.

——— . 1990. *N. A. Vasíliev e a Lógica Paraconsistente* [in Portuguese]. Coleção CLE 7. Campinas (SP): Centro de Lógica, Epistemologia e História da Ciência, Universidade Estadual de Campinas.

Bazhanov, V. A. 2007. *Istoriya logiki v Rossii i SSSR [History of Logic in Russia and the USSR]* [in Russian]. Moskva [Moscow]: Kanon+.

——— . 2025. "Predisloviye ko vtoromu, rasshirennomu izdaniyu [Preface to the Second Expanded Edition]" [in Russian]. In *Voobrazhayemaya logika. Izbrannyye trudy [Imaginary Logic. Selected Works]*, by N. A. Vasiliev, ed. by V. A. Bazhanov, 5–11. Moskva [Moscow]: Kanon+.

Church, A. A. 1936. "A Bibliography of Symbolic Logic." *The Journal of Symbolic Logic* 1 (4): 121–123.

Comey, D. D. 1965. "V. A. Smirnov. Logičéskié vzglády N. A. Vasil'éva [The Logical Views of N. A. Vasil'év]." *The Journal of Symbolic Logic* 30 (3): 368–370.

Costa, N. C. A. da. 1993. *Sistemas Formais Inconsistentes [Inconsistent Formal Systems]* [in Portuguese]. Curitiba: Editora UFPR.

D'Ottaviano, I. M. 1990. "Prefácio" [in Portuguese]. In *N. A. Vasíliev e a Lógica Paraconsistente*, by A. I. Arruda, XI–XVI. Coleção CLE 7. Campinas (SP): Centro de Lógica, Epistemologia e História da Ciência, Universidade Estadual de Campinas.

Kline, G. L. 1965. "Vasiliev and the Development of Many-Valued Logic." In *Contributions to Logic and Methodology in Honor of J. M. Bocheński*, ed. by A.-T. Tymieniecka, 315–326. Amsterdam: North-Holland.

Mal'tsev, A. I. 1976. *Izbrannyye trudy [Selected Works]* [in Russian]. Vol. 1. Moskva [Moscow]: Nauka.

Marcos, J. 2005. *Logics of Formal Inconsistency*. Campinas: CLE-Unicamp.

Markin, V., and D. Zaitsev, eds. 2017. *The Logical Legacy of Nikolai Vasiliev and Modern Logic*. Cham: Springer.

Paulhan, F. 1911. *La Logique de la contradiction* [in French]. Paris: Felix Alxan.

Poincaré, H. 2013. *Dernières pensées* [in French]. Paris: Flammarion.

Smirnov, V. A. 1989a. "Logicheskiye idei N. A. Vasil'yeva i sovremennaya logika [Logical Ideas of N. A. Vasiliev and Modern Logic]" [in Russian]. In *Voobrazhayemaya logika. Izbrannyye trudy [Imaginary Logic. Selected Works]*, by N. A. Vasiliev, ed. by V. A. Smirnov, 229–259. Moskva [Moscow]: Nauka.

——— . 1989b. "The Logical Ideas of N. A. Vasiliev and Modern Logic." In *Logic, Methodology and Philosophy of Science VIII*, ed. by J. E. Fenstad, I. T. Frolov, and R. Hilpinen, 625–640. New York: Elsevier.

——— . 1993. "Mnogomernyye logiki [Multidimensional Logic]" [in Russian]. *Logicheskiye issledovaniya [Logical Investigations]* 2:259–278.

Vasiliev, N. A. 1989. *Voobrazhayemaya logika. Izbrannyye trudy [Imaginary Logic. Selected Works]* [in Russian]. Ed. by V. A. Smirnov. Moskva [Moscow]: Nauka.

——— . 2025. *Voobrazhayemaya logika. Izbrannyye trudy [Imaginary Logic. Selected Works]* [in Russian]. Ed. by V. A. Bazhanov. Moskva [Moscow]: Kanon+.

Windelband, W. von, and A. Ruge, eds. 1912. *Die Prinzipien der Logik, Encyclopädie des philosophischen Wissenschfaten, Erster Band: Logik* [in German]. Tübingen: Verlag von Y. C. Mohr.

Da Mata J. V. T. [Да Мата Ж. В. Т.] The Development of Vasiliev's Ideas and Paraconsistent Logic in Russia and Outside [Развитие идей Васильева и паранепротиворечивой логики в России и за ее пределами] : A Review of the Second Russian Edition of Vasiliev's "Imaginary Logic" [рецензия на второе русское издание «Воображаемой логики» Васильева] // Философия. Журнал Высшей школы экономики. — 2025. — Т. 9, № 4. — С. 267–274.

Жозе Вериссимо Теишейра да Мата
профессор
Бразильское логическое сообщество (Бразилиа)

Развитие идей Васильева и паранепротиворечивой логики в России и за ее пределами

рецензия на второе русское издание «Воображаемой логики» Васильева

Васильев Н. А. Воображаемая логика. Избранные труды / под ред. В. А. Бажанова. — М. : Канон+, 2025.

DOI: 10.17323/2587-8719-2025-4-267-274.