Alexander Mikhailovsky and Elena Seredkina*

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

The Russian and Chinese Cases of AI Ethics

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 pro-
grams 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 re-
searchers 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 vul-
nerable groups. Nearly 80% of the analyzed documents contained references
to diversity and inclusion as ethical principles of AI, yet only about 20% pro-
vided 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 recommen-
dations predominate, while progress toward a formal treaty is repeatedly
obstructed by cultural, political, and sovereignty-related divides. A char-
acteristic 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, multistakeholder 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 nonliberal 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/paper s/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-intelli gence/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.

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

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.