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.