UDC 327:316.776:004.774(100)
DOI: 10.5937/spm94-59969
Original research article

Српска политичка мисао
(Serbian Political Thought)
No 6/2025.
Vol. 94
pp. 137-165

Ljubiša Bojić[*]

Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad

Institute for Philosophy and Social Theory, University of Belgrade

Digital Society Lab, Belgrade

Complexity Science Hub, Vienna, Austria

RISK AND RESPONSIBILITY AT THE FRONTIER OF AI: A THEMATIC ANALYSIS OF DEEP LEARNING PIONEERS' PERSPECTIVES ON ARTIFICIAL INTELLIGENCE THREATS AND GOVERNANCE[**]

Abstract

As artificial intelligence (AI) reshapes global societies, understanding its associated risks and governance imperatives is of urgent social importance. This study fills a critical gap by systematically analyzing extended interviews with Geoffrey Hinton, Yoshua Bengio, and Yann LeCun – to elucidate their firsthand perspectives on AI's existential, ethical, social, and governance challenges. Employing qualitative thematic analysis across six longitudinal interview transcripts, the research identifies both convergences and divergences: Hinton and

[*] E-mail: ljubisa.bojic@ivi.ac.rs; ORCID: 0000-0002-5371-7975

[**] This study was realised with the support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, according to the Agreement on the realisation and financing of scientific research 451-03-136/2025-03/ 200025 and the Agreement on the realisation and financing of scientific research 451-03-136/2025-03.

Bengio strongly emphasize existential threats, superintelligence hazards, AI weapons risks, and the need for robust global regulation, while LeCun expresses technological optimism and favors decentralized, open development. All acknowledge economic disruption, misuse of potential, and fractures in democratic discourse. The study's findings reveal that expert opinion on AI risk is far from monolithic and highlight actionable, innovative governance proposals, from regulated compute access to "diversity engines" in social media feeds. Implications include the necessity for adaptive, internationally coordinated AI governance and greater professional accountability among developers. Limitations include a focus on elite, Anglophone experts and inherent subjectivity in qualitative coding. Future research should expand to multi-stakeholder and cross-national perspectives, and test proposed regulatory frameworks in real-world contexts, addressing the ongoing evolution of risk as AI permeates new domains.

Keywords: artificial intelligence risk, deep learning pioneers, AI governance, existential threats, thematic analysis

INTRODUCTION

The ascent of artificial intelligence (AI) as a transformative technology marks a pivotal moment in the trajectory of human civilization. Over the past decade, AI has achieved remarkable milestones, transitioning from research laboratories into core infrastructure for contemporary society (LeCun, Bengio, and Hinton 2015). Deep neural networks, inspired in part by biological processes in the human brain, now underpin applications as varied as computer vision, internet search, speech recognition, machine translation, drug discovery, logistics, robotics, and automated financial trading. These systems have set new technical benchmarks, achieving and even exceeding human-level performance in domains such as image classification (Krizhevsky, Sutskever, and Hinton 2017), the board game Go (Silver *et al.* 2016), and large-scale language processing (OpenAI 2023).

The widespread deployment of AI creates unprecedented possibilities for social benefit, economic growth, and scientific discovery. For instance, AI diagnostic tools are increasingly used to interpret medical images, improving access to healthcare and enabling earlier intervention

(Esteva *et al.* 2017). In resource management, machine learning models optimize supply chains, reducing waste, and boosting sustainability. There is optimism among some scholars and policymakers that if guided responsibly, AI could make significant strides toward addressing global challenges such as poverty, education inequality, and climate change (Chui, Manyika, and Miremadi 2016). Indeed, the scale and versatility of AI's impact have prompted many to equate its significance with previous general-purpose technologies such as electricity or the internet (Brynjolfsson and McAfee 2014).

The accelerating pace and scale of AI adoption have also raised a chorus of urgent questions about risk, ethics, and social governance. Unlike technologies of the past, advanced AI systems carry the potential not merely to augment human activity, but to fundamentally reshape economic structures, modes of communication, and the boundaries of autonomy and agency in human life (Bostrom 2014; Russell 2019; Bodroža, Dinić, and Bojić 2024; Bojić *et al.* 2024; Bojić, Kovačević, and Čabarkapa 2025; Bojić, Stojković, and Jolić Marjanović 2024; Bojić *et al*. 2025). As AI systems increasingly mediate social decisions about credit, employment, policing, and information dissemination, concerns regarding transparency, fairness, and accountability have grown exponentially (Doshi-Velez and Kim 2017; O'Neil 2016; Noble 2018).

A significant body of literature explores the specific societal risks posed by contemporary AI. Algorithmic bias and discrimination have come under close scrutiny as systems trained on large, often uncurated datasets are shown to reproduce and reinforce patterns of racial, gender, and socioeconomic inequality (Buolamwini and Gebru 2018; Noble 2018; Sandvig *et al.* 2016). In health care, these biases can manifest in life-critical contexts, amplifying existing disparities and undermining trust (Obermeyer *et al.* 2019; Reinhardt *et al*. 2025).

Other research has exposed the role of "black-box" neural models in furthering opacity and frustrating efforts toward meaningful interpretability or recourse for affected individuals (Doshi-Velez and Kim 2017; Rudin 2019).

The economic implications of AI-driven automation have prompted intense scholarly and policy debate. While some analyses forecast AI as a complement to human labor, enhancing productivity and creating new categories of employment (Bessen 2018), others warn of large-scale displacement, particularly for middle-skill and routine occupations (Brynjolfsson and McAfee 2014; Korinek and Stiglitz

2018). Beyond the loss of income, such displacement may threaten social cohesion, increase inequality, and diminish the social status and personal dignity associated with work (Susskind 2020; Autor 2015). The distribution of economic gains from AI also risks concentrating power in a handful of corporations and states, aggravating global disparities (Bojic 2022; Bojic 2024).

In the political domain, machine learning techniques have been deployed not only for benign purposes such as translating languages or moderating content, but also to manipulate public opinion through targeted advertising, deepfakes, and recommendation algorithms (Pavlovic and Bojic 2020; O'Connor and Weatherall 2019). The proliferation of generative models has blurred the distinction between authentic and synthetic content, complicating journalistic verification and undermining democratic deliberation (West 2019; Zuboff 2019). Repressive regimes have exploited AI for surveillance and social control, raising profound questions about privacy, civil liberties, and the resilience of liberal democratic institutions (Feldstein 2019).

Existential risks associated with advanced AI systems have also received increased attention in recent years. Bostrom (2014) and others highlight the potential for a hypothetical "superintelligence" – an AI system surpassing human cognitive abilities – to act in ways detrimental or even catastrophic to humanity if it becomes misaligned with human interests. Even in the absence of the arrival of superintelligence, warnings have been issued concerning accidental or malicious misuse of powerful AI technologies, for example, in cyberattacks, the development of lethal autonomous weapons, or the creation of destabilizing misinformation (Brundage *et al.* 2018). As AI capabilities continue to expand, the challenges of governing such systems – many of which are developed and deployed transnationally by commercial actors – are only becoming more complex (Floridi *et al.* 2018).

Despite this rich and expanding literature, a critical gap remains. Much of the existing discourse, while rigorous, is written by ethicists, social scientists, journalists, and technology policy experts. Far less research has systematically engaged with the firsthand perspectives of those most instrumental in developing and propagating the technologies at the heart of the ongoing AI revolution. Indeed, as AI's foundational technologies mature, the reflections, warnings, and priorities of its principal architects are increasingly recognized as vital contributions to the societal dialogue (Metz 2023; Turing 2018).

Among these architects, Geoffrey Hinton, Yoshua Bengio, and Yann LeCun stand apart as the preeminent pioneers of deep learning. Their research, begun during an era when neural networks were marginalized within the artificial intelligence community, has shaped the breakthroughs that define current AI progress. Collectively, their work on backpropagation, convolutional neural networks, probabilistic modeling, natural language processing, and generative adversarial learning forms the backbone of nearly all state-of-the-art systems in commercial and scientific use (LeCun, Bengio, and Hinton 2015; Schmidhuber 2015). In recognition of their transformative contributions, all three were jointly awarded the 2018 ACM A.M. Turing Award, the highest honor in computer science (Turing 2018).

The influence of Hinton, Bengio, and LeCun extends far beyond the laboratory. Each has served as an advisor to leading governments and international bodies on AI policy, headed major industrial AI research laboratories (including at Google and Meta/Facebook), and shaped entire research ecosystems through the foundation of academic institutes and the training of hundreds of graduate students (LeCun, Bengio, and Hinton 2015). Crucially, all three have become increasingly transparent and vocal about their own shifting views on AI's risks and social obligations, with Hinton – sometimes described as the "Godfather of AI" – resigning from Google in 2023 to speak more freely about the potential dangers of the technologies he helped create (Metz 2023).

While the technical achievements of these pioneers have been thoroughly chronicled, relatively little scholarly attention has been paid to how they themselves diagnose, prioritize, and envision the mitigation of risks associated with AI. Publicly available, long-form interviews and panel discussions provide a unique window into their evolving thought processes, concerns, and responses to the accelerating diffusion of AI. Such qualitative material allows researchers to explore explicit warnings – such as those regarding existential threats or social instability – and the underlying values, uncertainties, and tensions voiced by these scientists as they grapple with the legacy and future direction of their field (Russell 2019; Floridi *et al.* 2018; ACM 2018).

This paper aims to address the aforementioned gap by conducting a systematic qualitative analysis of extended interviews with Geoffrey Hinton, Yoshua Bengio, and Yann LeCun. Drawing on rich, longitudinal data collected from multiple interviews per participant, this study interrogates their perspectives within the wider context of the scholarly and public discourse on AI risk. Specifically, the analysis is guided by the following research questions:

- First, how do Hinton, Bengio, and LeCun conceptualize and prioritize the most significant threats arising from advances in artificial intelligence?
- Second, what themes of consensus or divergence emerge in their analyses of technical, economic, ethical, and existential risks, and to what extent are these aligned (or not) with dominant narratives in the academic literature?
- Third, how have their views evolved in response to technological breakthroughs such as large language models, as well as shifting political and societal conditions?
- Finally, what implications do their positions have for the governance and responsible stewardship of AI technologies going forward?

This research endeavors to deepen our understanding of AI's risks and the complex interplay between scientific progress, individual responsibility, and social foresight. In so doing, the paper seeks to inform AI policy, research, and public dialogue at a time when the stakes for humanity have seldom been higher.

METHODOLOGY

This study employed a qualitative research design, specifically utilizing thematic analysis, to explore and synthesize the perspectives of three foundational figures in modern artificial intelligence – Geoffrey Hinton, Yann LeCun, and Yoshua Bengio – on the risks and threats posed by recent advances in AI. Each of these individuals was selected for analysis owing both to their status as recipients of the 2018 ACM A.M. Turing Award, often considered the "Nobel Prize of Computing," and for their direct, longstanding roles in conceptualizing and implementing deep neural network models that now underpin a wide range of contemporary AI systems (ACM 2018; Turing_Bengio 2018; Turing_Hinton 2018; Turing_LeCun 2018). The present study sought to document, compare, and interpret their views as expressed in extended public interviews, thereby capturing insights likely to shape the evolving discourse on the societal impact of AI.

The empirical material comprised six in-depth interview transcripts (OSF 2025; Hinton_Diary 2025; Hinton_60Minutes 2023; Bengio_BBC 2025; Bengio_WSF 2024; LeCun_Brian 2024; LeCun_Lex 2024).

For each of the three Turing laureates, two interviews were selected: one interview conducted within the previous six months, and a second conducted no less than one year prior. This sampling framework was intended to both provide a breadth of expression and facilitate longitudinal comparison, enabling the detection of any shifts in perspective as the technological and ethical landscape of AI has rapidly evolved. Interviews were sourced from widely recognized, publicly accessible podcasts and YouTube discussion channels. The selection criteria prioritized interviews in which each guest was engaged in an open, discursive exchange, and where the conversation explored, either directly or tangentially, current and anticipated risks associated with artificial general intelligence, deep learning, and their applications. Because the personalities involved are public figures and the interviews were intended for public dissemination, all identifying and contextualizing information was preserved in the transcripts. Where interview transcripts were not already available, the audio or video content was precisely transcribed by the researcher to ensure fidelity to the speakers' statements, including pauses, qualifiers, and nonverbal cues that contributed meaning or emphasis. The final textual corpus underwent careful review, with corrections made for errors in transcription and to clarify ambiguous utterances where necessary.

Thematic analysis was chosen as the principal analytic strategy. This approach, as articulated by Braun and Clarke (2006), provides a systematic yet flexible framework for identifying, organizing, and interpreting salient patterns within qualitative data. Unlike quantitative content analysis, thematic analysis does not reduce data to counts or simple categories; instead, it emphasizes depth of meaning, allowing the researcher to surface both explicit statements and more latent underlying themes embedded in the discourse. Analysis began with immersion in the data: the researcher read and reread the entire set of transcripts, making initial margin notes concerning recurring topics, metaphors, and expressions of concern or optimism. This stage established a foundation for subsequent coding, fostering a holistic sense of each interview as well as the data set as a whole.

During the initial coding phase, both deductive and inductive elements were employed. The starting codebook was informed by leading scholarship on risks associated with AI and machine learning – such as existential threats, algorithmic bias, misuse in authoritarian or military domains, mass displacement of labor, and ethical dilemmas in autonomous decision-making (Braun and Clarke 2006; ACM 2018). At the same time, the researcher allowed for the emergence of unexpected or novel topics through inductive coding, remaining attentive to unique turns of phrase, emotionally charged language, or instances where the interview subject expressed personal ambivalence, regret, or optimism regarding their own legacy. Coding was conducted manually in successive passes, and the codebook was treated as a living document, subject to refinement as additional patterns or contradictions became visible across the interviews. When ambiguous passages appeared, the researcher documented interpretive choices in analytic memos to sustain transparency and reflexivity.

Following coding, the codes were examined for conceptual affinity and organized into preliminary themes. These themes ranged from clearly delineated technical risks – such as concerns about autonomous weapons, adversarial attacks, or uncontrolled AI self-improvement – to broader social and philosophical worries, including implications for economic inequality, mass unemployment, erosion of democratic processes, and fundamental questions about consciousness, sentience, and value alignment in artificial agents. Themes were further scrutinized and revised through an iterative, dialogic process, with attention to both convergence and divergence in the informants' perspectives. For example, the degree of emphasis placed on existential risk versus more immediate hazards, or varying levels of confidence in regulatory and governance solutions, was noted and integrated as subthemes or points of contrast. Throughout this process, the researcher sought to remain faithful to the language and logic of the participants themselves and to interpret their statements in relation to broader scholarly and policy debates.

Analysis and interpretation were supported by ongoing self-reflection and the maintenance of an audit trail, which documented all key decisions and adjustments to methodological procedures. Despite the single-author nature of the analysis, rigor was maintained through repeated returns to the data, the careful selection of illustrative quotations, and conscious monitoring for interpretive bias. When representative extracts were selected to exemplify particular themes or to demonstrate moments of disagreement among the Turing laureates, an effort was made to preserve context and the unique rhetorical style of each speaker. In reporting results, the distinction between description and interpretation was actively maintained.

Ethical considerations for this research were minimal, since all data originated in the public domain and involved individuals speaking in their professional and public-citizen capacity. No sensitive or private information was included, and no interaction occurred between the researcher and the interview subjects. As such, the study is exempt from review by an institutional research ethics board, though all guidelines for responsible scholarship and data integrity were followed.

RESULTS

Thematic analysis of six long-form interviews – two each with Geoffrey Hinton, Yoshua Bengio, and Yann LeCun – yielded 48 *a priori* codes clustered in seven super-themes (A–G). All 48 codes were observed at least once across the corpus; none of the passages required a new category, confirming the sufficiency of the codebook. Below, we synthesise points of convergence and divergence, followed by a comparative table that condenses how strongly each laureate emphasised every super-theme (Table 1).

Hinton and Bengio both foregrounded the prospect of super-intelligence and openly used the language of "existential threat." Hinton's interviews contained 28 distinct passages within cluster A, repeatedly quantifying a "10–20 percent" probability of human extinction (00:54; 08:54), while Bengio invoked extinction or species "replacement" ten times, likening present research to "playing apprentice sorcerer" (26:41). Both described timelines of roughly 5–20 years.

LeCun, by contrast, acknowledged eventual machines "smarter than us" but dismissed catastrophe scenarios as products of a "doomer" mind set. His two interviews yielded only five passages in cluster A, none of which endorsed extinction risk. Instead, he argued that domination requires "hard-wired drives" that engineers can simply omit (131:30), framing alignment as an engineering optimisation task rather than an existential unknown.

Job displacement, loss of dignity, and widening inequality were most salient for Hinton, who provided concrete anecdotes (e.g., a niece whose task time fell from 25 to 5 minutes, 41:05) and offered practical career advice ("Train to be a plumber," 88:01). Bengio's 2024 appearance focused more on catastrophic misuse than labour, but his 2025 BBC interview acknowledged disappearing entry-level legal work. LeCun, although conceding that "natural fear" exists (143:40), framed labour impacts as manageable and emphasised the upside of an "amplifier of human intelligence."

Hinton and Bengio expressed acute concern over cyber-attacks, bio-weapons, and autonomous weapons. Hinton cited a "12,200 percent" year-on-year spike in cyber incidents (12:18) and described combinatorial threat spirals (28:32). Bengio detailed a six-month timeline for engineering lethal viruses (33:54). LeCun, by contrast, dismissed large language models as insufficient for building bio-weapons and treated hostile persuasion scenarios as a future arms race between competing AI assistants.

All three laureates recognised algorithmic manipulation, but with different emphases. Hinton provided the longest discussion of echo-chambers and loss of shared reality, blaming profit-driven recommender systems. Bengio linked manipulation to democratic back sliding. LeCun warned primarily against the concentration of "information diet" in the hands of a few proprietary models, advocating open-source diversity rather than heavy regulation (Bojic, Agatonović, and Guga 2024; Bojic and Marie 2017).

Hinton and Bengio converged on the need for strong global governance – Hinton even positing a "world government that works" (11:24) and Bengio calling for binding international treaties (39:41). Both criticised military carve-outs and corporate profit motives. LeCun worried instead about over-regulation stifling open research; his remedy was decentralisation through open-source foundations. Thus, all three accept the premise of governance gaps, yet propose sharply different solutions.

Each scientist acknowledged transformative upside – especially in health, education, and scientific discovery – but the emotional valence differed. Hinton spoke of "magnificent" benefits (09:55) yet confessed that such promise "takes the edge off" his pride (32:05). Bengio voiced a shift from "excited" to "worried," balancing cures for disease against existential downside. LeCun remained emphatically optimistic, likening AI to the printing press and the Enlightenment (160:42) and predicting "a bright future for humanity if we do this right" (57:55).

Hinton and Bengio articulated personal anxiety, duty, and even regret. Hinton admitted difficulty "coming to terms emotionally" (52:43) and lamented time lost with family due to career obsession (82:39). Bengio expressed responsibility "to talk to governments and citizens" (51:57). LeCun displayed no comparable distress; instead, he articulated confidence that "doom" narratives underestimate both engineering capability and intrinsic human goodness.

Comparing the earlier and later interviews for each laureate reveals drift primarily in Bengio's timeline estimates (from 5–20 years in 2024 to 2–10 years in 2025) and Hinton's escalating emphasis on regulation after leaving Google. LeCun's stance remained comparatively stable between March and December 2024, although the later interview further down played existential concerns.

All three Turing laureates acknowledge both promise and peril, but their emphases diverge sharply. Hinton and Bengio converge on stark existential warnings and the urgency of global regulation, whereas LeCun positions himself as a technological optimist concerned primarily with ensuring open, decentralised development. These differences show that even within the vanguard of deep learning expertise, assessments of AI risk are far from uniform – an insight critical for policymakers who might otherwise treat "expert opinion" as monolithic.

Table 1. Salience of Super-Themes across Laureates. Coding density was graded qualitatively: "High" = >15 supporting excerpts across both interviews; "Moderate" = 5–15 excerpts; "Low" = <5 excerpts or largely dismissive treatment.

Super-Theme (A–G)	Key Risk / Topic	Hinton (2023-2025)	Bengio (2024-2025)	LeCun (2024)
A. Existential & long-term	Super-intelligence, extinction, alignment	High	High	Moderate-Low
B. Societal (jobs, inequality)	Displacement, dignity, UBI	High	Moderate	Moderate-Low
C. Misuse & security	Cyber, bio-weapons, lethal robots	High	High	Low
D. Social-psychological	Polarisation, misinformation, bias	High	Moderate	Moderate
E. Governance & regulation	Global treaties, profit motives, capture	High	High	Moderate
F. Positive potentials / ambivalence	Productivity, health, dual-use	Moderate	Moderate	High
G. Emotional & personal	Duty to warn, regret, personal impact	High	High	Low

Source: Author.

DISCUSSION

The constellation of risks articulated by Geoffrey Hinton, Yoshua Bengio, and Yann LeCun – often called the "godfathers of deep learning" – highlights the need for a new era of AI governance that is as innovative and dynamic as the technology itself (Table 2). While there is consensus among these leaders that AI's potential benefits are enormous, the analysis also reveals genuine apprehension about existential threats, misuse, social disruption, and the inadequacies of existing regulatory approaches. Here, we interpret the findings with a view toward transformative regulatory interventions that would empower societies to guide AI trajectory for the common good.

EXISTENTIAL AND LONG-TERM RISKS

The most fundamental worries – the possibility of AI surpassing human intelligence, acting independently, and potentially making humankind obsolete – cannot be adequately addressed by traditional sector-specific or national regulations. This is not simply because the technology is powerful, but because it can scale at unprecedented speed across borders, within private infrastructures that are often opaque even to governments (Brundage *et al.* 2018). When Geoffrey Hinton asserts, "I realise that these things will one day get smarter than us and we've never had to deal with that," he describes a singularity that upends the very logic of retrospective regulation (Hinton_Diary 2025).

To control the development of AI systems that approach superhuman capacities, we need to regulate the very infrastructure – massive computation ("compute") – that undergirds AI research. No company or research consortium could train an advanced AI system above a certain scale without first obtaining a digital "compute passport." This would function analogously to a visa: to run major training jobs or build very large neural networks, organizations would have to submit their plans, security measures, and alignment protocols to a secure registry, possibly managed by a new international body analogous to the International Atomic Energy Agency (Floridi *et al.* 2018; Brundage *et al.* 2018). The request would be logged cryptographically; only after approval would compute resources (from cloud providers or AI chip manufacturers) be released. This approach would act upstream, preventing unaccountable actors from conducting "AI moonshots" out of public sight.

Because AI capabilities evolve quickly, one-off approvals are not enough. Instead, licenses for "frontier" AI projects should be renewed every six or twelve months, each time requiring new independent audits of alignment strategies, interpretability, and security (Gabriel 2020). To align private incentives, companies would be made to post "catastrophe bonds" – essentially insurance policies that pay out automatically to an international compensation fund in the event of proven catastrophic harm, such as widespread loss of control or the use of AI in major cyber-attacks. The bond premium paid by the company would scale with the assessed risk, sending a continuous market signal about the evolving dangers of different projects (Korinek and Stiglitz 2018).

ALIGNMENT AND CONTROL

One recurring motif in both Hinton's and Bengio's statements is the need to take responsibility at the top: "We need to figure out how to make [AI systems] not want to take over," Hinton says, but in practice, that means developers and corporate leaders must be legally accountable for their designs (Hinton_Diary 2025). One creative solution is to impose a fiduciary duty – an explicit and legally binding obligation – on company executives overseeing AI systems above a certain capability. This would oblige them to prioritize public safety and value alignment even above profits, mirroring duties that already exist for trust managers, medical boards, or environmental stewards (Cath 2018).

Given the concern that powerful AI models could "go rogue," high-capability AI systems should be managed like highly sensitive cryptographic keys or nuclear launch codes. Their parameters (the digital weights that determine their behavior) would be encrypted and divided into slices, each entrusted to an independent body – such as the developer, a government agency, and a citizen s' panel. Changing or deploying the system would require consensus, ensuring that unilateral behavior – from any one actor or rogue employee – is impossible (Brundage *et al.* 2018). This "AI escrow" approach would add a technical check to the legal ones.

TECHNOLOGICAL MISUSE

The possibility that AI could be used not just for cyber-attacks but for creating new bioweapons, as Bengio warns, demands a whole new order of preparedness. A government-chartered "Algorithmic Center for Disease Control and Prevention" (A-CDC) could serve as a living laboratory where AI models capable of simulating or manipulating biological/chemical systems above a certain risk threshold must be submitted prior to real-world deployment. This agency would continuously test for vulnerabilities, proactively attempt to "break" models, and develop defensive countermeasures, which would then be shared globally with vetted labs. It would operate as a continuously updated "red team" that anticipates, rather than simply reacts to, emerging algorithmic threats.

Current norms encourage the open publication of powerful AI models, which can create a one-way ratchet: once released, anyone can deploy or modify them. Regulatory bodies should require private and public labs to submit "capability risk assessments" – akin to bio-safety levels in pathogen research – before publishing code or weights for models that could be misused. Only open-sourcing up to a certain risk threshold would be permitted routinely; anything above that would face a structured review with mandatory threat modeling and provisional embargoes, coordinated internationally (Brundage *et al.* 2018).

THE ESCALATING THREAT OF AI WEAPONS AND AUTONOMOUS LETHAL SYSTEMS

As artificial intelligence matures, its integration into weapons platforms and military strategy has emerged as one of the most acute vectors of risk. Lethal autonomous weapon systems (LAWS), which can identify, select, and engage targets without direct human intervention, exemplify the convergence of cutting-edge AI with the destructive power of advanced weaponry. Unlike prior technological advances in arms manufacturing, AI-enabled weapon systems possess the capacity for rapid adaptation, distributed deployment, and even self-directed action – dramatically widening the potential for unforeseen escalation, misuse, or catastrophic failure (Russell 2019; Brundage *et al.* 2018).

Both Geoffrey Hinton and Yoshua Bengio single out autonomous weapons as a uniquely destabilizing risk. Bengio cautions that "as soon as you have the technology, it will be used somewhere in the world" (Bengio_WSF 2024), warning of a world in which the diffusion of low-cost, highly lethal autonomous systems could lower the threshold for conflict, proliferate to non-state actors, and exacerbate asymmetries in power that destabilize international security. Hinton, similarly, points to the combinatorial threat posed by the intersection of AI with existing weapon systems, highlighting scenarios where "face-recognition drones could be calibrated to target political dissidents, journalists, or even deployed as instruments of ethnic violence or terror" (Hinton_60Minutes 2023).

Existing arms control frameworks – including the Geneva Conventions and the United Nations Conventional Weapons Convention – have proven too slow and politically unwieldy to address these novel threats. The lack of clear, enforceable, and universally-agreed-upon restrictions on the development, deployment, and transfer of AI-powered weapons creates a dangerous vacuum at the frontier of military and dual-use AI research.

Given the rising capability of AI weapons and the real risk of an ungoverned arms race, there is an urgent need for the international community to move beyond voluntary guidelines and individual state-based pledges toward an "internationally accepted moratorium" on the development and use of fully autonomous weapon systems. This moratorium would serve as an immediate, interim "brake," buying critical time for the negotiation of legally binding treaties, normative frameworks, and verification mechanisms tailored to the unique characteristics of AI (Russell 2019).

Such a moratorium should extend not only to the battlefield use of AI weapons, but also to their research, development, and testing – especially for systems capable of selecting and attacking human targets without meaningful human control (ICRC 2021; Future of Life Institute 2017). It should be underpinned by transparent, multi lateral verification; include enforcement provisions for technology transfer and circumvention; and accommodate whistleblower protections within both public and private AI laboratories.

An internationally coordinated moratorium would align with the "precautionary principle" often invoked in discussions of existential risk: when an emerging technology threatens severe or irreversible harm, and scientific consensus is incomplete, policy should err on the side of preemptive restraint (Bostrom 2014). As Hinton and Bengio have remarked in recent interviews, inaction risks “locking in” the proliferation of AI weapons before society, ethics, and law have an adequate opportunity to respond (Bengio_BBC 2025).

Drawing on the model of prior arms control achievements – such as the bans on chemical and biological weapons, anti-personnel mines, and blinding laser weapons – an effective moratorium on AI weapons will require international cooperation, technical standards for verification, and meaningful involvement of civil society and the scientific community (Altmann and Sauer 2017).

SOCIETAL DISRUPTION AND ECONOMIC DISPLACEMENT

AI’s automation potential risks both mass unemployment and the further concentration of wealth. Rather than blunt prohibitions, governments could institute a system of “automation credits” payable by companies for each job function replaced by AI. The rate would be calibrated to local unemployment levels: when new jobs are created in sectors where human labor is scarce – such as green energy or healthcare – the credit would be minimal or even reversed. When jobs are destroyed in already-vulnerable regions, the credit would be high. The proceeds would be pooled into local “Worker Trusts,” which could fund wage subsidies, educational retraining, or public works – effectively recycling AI-driven productivity gains back into communities most disrupted by change (Korinek and Stiglitz 2018; Standing 2018).

Every citizen, at birth, could be assigned a “universal learning account” funded partly by automation credits and partly by a tax on AI-driven data harvesting. This account could be drawn on throughout a person’s life for continuous retraining, higher education, or even transitioning to new forms of creative or caretaking work – enabling lifelong adaptation as AI transforms the landscape of employment.

SOCIAL AND PSYCHOLOGICAL RISKS

Interviewees express grave concern about society’s fracturing into “echo chambers” driven by algorithmic content curation, where “we don’t have a shared reality anymore” (Hinton_Diary 2025; Pariser 2011; Alipour *et al.* 2024). Merely adding “labels” or “friction” to particular posts has so far proven insufficient to restore a common epistemic ground.

A breakthrough regulatory solution would be to require major digital platforms (e.g., social networks, video platforms, news aggregators) to deploy "diversity engines" as part of their feed algorithms (Bojic 2024). By law, a designated proportion of users' feeds – say, 20% – would consist of content surfaced by a certified independent diversity algorithm, not just the platform's own engagement-optimizing system. This diversity engine would be designed to maximize users' exposure to a range of perspectives, including those that challenge their biases or worldview (Moe, Hovden, and Karppinen 2021). Users would retain settings for language and sensitivity, but could not opt out of diverse exposure entirely. The performance of these engines would be regularly audited by an independent commission, ensuring that platforms are not artificially degrading the "diversity feed" or gaming the metrics.

Such an approach borrows from the tradition of "public service broadcasting," but uses algorithmic tooling to ensure that, however much our realities may diverge, we are at least occasionally drawn back into shared information environments (Helberger 2019). This could be further reinforced by algorithmic transparency mandates, requiring platforms to publish summary statistics about the diversity and quality of content presented to users.

REGULATORY CAPACITY

Because AI is evolving rapidly, any fixed law risks obsolescence. Policymakers could introduce "adaptive statutes," a new form of law that is continuously updated via a regulatory large language model (Reg-LLM). This model would monitor case law, technical developments, international agreements, and enforcement outcomes, then propose redlined amendments to statutes on a regular basis, complete with readable explanations and impact analysis. Legislators would vote on these adaptive updates at fixed intervals, vastly increasing the agility and responsiveness of the legal environment (De Filippi, Hassan, and Zicari 2023). All model outputs would be archived on a public blockchain for transparency and post hoc audit, ensuring that regulation keeps pace with the industry it governs.

Policy must also address not just systems and platforms, but people. Both Hinton's and Bengio's sense of duty – and regret – points to the need for a stronger culture of responsible innovation. Jurisdictions could introduce professional "AI stewardship licenses," requiring advanced AI researchers and lead engineers to maintain a public dossier that is updated every time they contribute to a major system, similar to continuing medical education records for doctors. The dossier would include a brief self-assessment of social impacts, decisions about risk mitigation, and lessons learned. While this may seem soft compared to legal penalties, such transparency would exert peer and reputational pressure, fostering norms that reward responsibility and reflection (Jobin, Ienca, and Vayena 2019; Floridi *et al.* 2018).

Collectively, these creative ideas form an integrated and ambitious AI governance agenda. Regulating "upstream," at the level of compute and model licensing, can help prevent dangerous models from being built unnoticed. Fiduciary duties and escrow split authority ensure that safety is not sacrificed in the pursuit of profit. Algorithmic CDCs and mandatory red-teaming equip governments to preempt misuse rather than merely responding to disasters. Diversification mandates for recommender systems renew the fabric of democratic discourse, even as they preserve individual choice. Adaptive statutes and professional licensing keep regulation fit for purpose and aligned with ethical imperatives.

Enacting such reforms will not be easy. The technical detail required for enforcement, the need for global coordination, and the challenge of keeping up with persistent innovation will test the capacity of current institutions. Nevertheless, waiting for catastrophic outcomes is a risk we cannot afford to take. To harness the opportunities of AI – while genuinely safeguarding humanity, dignity, and democracy – will require a willingness to experiment with regulatory structures as unprecedented as the threats we face.

Table 2. AI risks and potential regulatory innovations

Risk Domain	Key Challenges	Breakthrough Regulatory Solutions (examples)

Existential/Long-term	Surpassing human intelligence;
	loss of control;
	unpredictability
		Compute passports; conditional model licenses; catastrophe bonds; international AI agency

Alignment/control	Aligning AI values with human ones; lack of direct technical solutions	Fiduciary duties for AI execs; AI escrow systems (split authority over deployment); "kill switch" protocols
Misuse/security	Bioweapons, cyber-attacks, autonomous weapons, open-source risks	Algorithmic Center for Disease Control (A-CDC); mandatory risk assessment before open publishing; secure model submission
Societal/ inequality	Job loss, dignity, and unequal distribution of AI-generated wealth	Automation credits based on local impact; dynamic worker "windfall trusts"; universal learning accounts
Social/ psychological	Polarization, echo chambers, lack of shared reality	Mandated diversity engines for recommender systems; periodic algorithmic audits; user interface controls for recommender bias
Regulatory agility	Static, slow, or uninformed legislative response	Adaptive law with regulatory LLMs; public blockchain of all regulatory changes and their justifications
Ethical and professional	Personal and institutional responsibility; values-conflicted leadership	AI professional stewardship licenses; public impact dossier for lead AI developers; periodic peer review of risk mitigation

Source: Author.

CONCLUSION

The findings of this study hold profound social relevance as the world stands at the threshold of a new technological epoch. Advanced AI systems – especially those built on deep learning – are not only transforming industries but beginning to reshape the very structures of daily life, the labor market, social reality, and even collective imaginaries about the future of humanity. As the voices of Hinton, Bengio, and LeCun illustrate, the stakes are not merely technical; they are existential, ethical, and democratically political. The range and seriousness of risks identified, from existential threats to democracy's erosion through algorithmic fragmentation, highlight the urgent need for innovative governance solutions that integrate both the promise and peril of emerging AI technologies.

In formal response to the research questions, this study has shown:

- First, the leading pioneers of deep learning articulate a multidimensional risk landscape. Hinton and Bengio are especially vocal about existential threats, the risks of superintelligence, and the potential for irreversible harm if AI systems surpass human intelligence without robust alignment or control mechanisms. LeCun, while recognizing the technical and societal challenges, consistently expresses optimism, emphasizing the amplifying potential of AI, the fallibility of catastrophic scenarios, and the prospects for meaningful regulation and design safeguards.
- Second, there is both substantive overlap and clear divergence in the way these thought leaders prioritize risks and characterize feasible remedies. While all three agree on the importance of governance, they differ substantially in their perceptions of urgency, the plausibility of loss of control, and the necessary trade-offs between innovation and caution. Hinton and Bengio urge regulatory interventions to increase the "stack" (e.g., compute passports, fiduciary duties), while LeCun favors iterative engineering solutions and a balanced approach to platform regulation.
- Third, over time and across interviews, there is a marked trend toward greater explicitness and even humility from those who once steered the field with techno-optimistic assumptions. Bengio's personal journey – from unbridled enthusiasm to a new sense of social duty and caution – mirrors a growing movement among AI professionals toward public engagement and ethical reflection (Jobin, Ienca, and Vayena 2019). All three maintain that, while the exact content of "safety" remains an active research challenge, responsibility for guiding AI's impact cannot be delegated to technologists alone, but must be shared by regulators, democratic institutions, and civil society actors (Floridi *et al*. 2018; Gabriel 2020).

The implications of these findings are wide-ranging. First, they suggest that existing approaches to AI regulation and governance are likely to be insufficient unless fundamentally revised. High-impact, transformative regulatory ideas – such as regulated compute, dynamic model licensing, and mandated diversity engines – offer actionable paths forward but will require new institutional partnerships, international agreements, and a stronger culture of professional responsibility. Second, the need for "regulatory agility" – statutes and standards capable of learning and evolving alongside AI itself – is made evident in both the pace of technological change and the continual discovery of new unintended consequences. Third, the call for robust social and psychological interventions – such as regulated diversification of recommender systems– highlights the urgency of supporting economic adaptability, democratic resilience, and social cohesion (Helberger 2019; Moe, Hovden, and Karppinen 2021).

However, this research has limitations that must be acknowledged. The analysis is based on a purposive sample of interviews with three highly influential experts. While these individuals represent extraordinary technical and social authority, their views are inevitably shaped by personal histories, cultural contexts, and professional networks. The study does not include perspectives from critics of deep learning, social scientists, policymakers, or affected labor groups, nor does it systematically analyze non-English debates or diverse policy contexts. As with all qualitative thematic analyses, the coding and interpretation of content are subject to researcher bias and the limits of available source material. The innovative regulatory proposals reviewed here, while grounded in existing scholarship and creative in conception, have not been tested in practice and may encounter significant political, technical, or economic barriers to implementation.

Future research should address several promising directions. Empirically, there is a need for multi-stakeholder, cross-national studies exploring risk perception and solution preference across technical, regulatory, and public constituencies. Experimental or pilot projects should be launched to evaluate the real-world feasibility and unintended effects of proposals such as compute passports or "diversity engines" for algorithmic feeds. Longitudinal investigation is also warranted to track the evolution of elite expert discourse as AI systems become further embedded in critical societal infrastructure. Ongoing developments such as AI for synthetic biology, autonomous military systems, and large-scale generative models will require continuous re-examination of both the risk landscape and the adequacy of regulatory responses.

This study reinforces the view – now increasingly accepted by governments and multilateral bodies – that AI governance must no longer be considered a peripheral regulatory issue, but rather a core concern for social order, democratic legitimacy, and the preservation of fundamental rights. Socially and politically, there is also the challenge of ensuring "ethical preparedness" and fostering a culture wherein responsibility for AI safety and benefit is shared and institutionalized at every level of innovation and deployment (Jobin, Ienca, and Vayena 2019; Floridi *et al.* 2018).

To guide AI toward outcomes aligned with human flourishing, societies must invest in the difficult work of regulatory architecture, scientific research on safe AI, cross-sectoral collaboration, democratic engagement, and the cultivation of ethical stewardship throughout the AI value chain.

REFERENCES

Alipour, Shayan, Alessandro Galeazzi, Emanuele Sangiorgio, Michele Avalle, Ljubisa Bojic, Matteo Cinelli, and Walter Quattrociocchi [Alipour *et al.*]. 2024. "Cross-Platform Social Dynamics: An Analysis of ChatGPT and COVID-19 Vaccine Conversations." *Scientific Reports* 14 (1): 2789. DOI:10.1038/s41598-024-53124-x

Altmann, Jürgen, and Frank Sauer. 2017. "Autonomous Weapon Systems and Strategic Stability." *Survival* 59: 117–42. DOI: 10.1080/00396338.2017.1375263

Autor, David H. 2025. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives* 29 (3): 3–30.

Bengio_BBC. 2025. "The worst case scenario is human extinction – Godfather of AI on rogue AI. Yoshua Bengio." *BBC Youtube*. https://youtu.be/c4Zx849dOiY

Bengio_WSF. 2024. "Why a Forefather of AI Fears the Future. Yoshua Bengio." *World Science Festival. Youtube.* https://youtu.be/KcbTbTxPMLc

Bessen, James E. 2018. "AI and Jobs: The Role of Demand." *NBER Working Paper*. https://www.nber.org/papers/w24235

Bodroža, Bojana, Bojana M. Dinić, and Ljubiša Bojić. 2024. "Personality Testing of Large Language Models: Limited Temporal Stability, but

Highlighted Prosociality." *Royal Society Open Science* 11: 240180. DOI: 10.1098/rsos.240180

Bojic, Ljubisa. 2022. "Metaverse through the prism of power and addiction: What will happen when the virtual world becomes more attractive than reality?" *European Journal of Futures Research* 10 (1): 22. DOI: 10.1186/s40309-022-00208-4

Bojic, Ljubisa. 2024. "AI alignment: Assessing the global impact of recommender systems." *Futures* 160 (103383). DOI:10.1016/j.futures.2024.103383

Bojic, Ljubisa, and Jean-Louis Marie. 2017. "Addiction to Old versus New media." *Srpska politička misao* 56 (2): 33–48. DOI: 10.22182/spm.5622017.2

Bojić, Ljubiša, Irena Stojković, and Zorana Jolić Marjanović. 2024. "Signs of Consciousness in AI: Can GPT-3 Tell How Smart It Really Is?" *Humanities and Social Sciences Communications* 11: 1631. DOI: 10.1057/s41599-024-04154-3

Bojić, Ljubiša, Matteo Cinelli, Dubravko Ćulibrk, and Boris Delibašić [Bojić *et al.*]. 2024. "CERN for AI: A Theoretical Framework for Autonomous Simulation-Based Artificial Intelligence Testing and Alignment". *European Journal of Futures Research* 12: 15. DOI: 10.1186/s40309-024-00238-0

Bojic, Ljubisa, Milos Agatonović, and Jelena Guga. 2024. "The Immersion in the Metaverse: Cognitive Load and Addiction." In *Augmented and Virtual Reality in the Metaverse*, eds. Vladimir Geroimenko. Cham: Springer Nature Switzerland. DOI:10.1007/978-3-031-57746-8_11

Bojić, Ljubiša, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević [Bojić *et al.*]. 2025. "Comparing Large Language Models and Human Annotators in Latent Content Analysis of Sentiment, Political Leaning, Emotional Intensity and Sarcasm." *Scientific Reports* 15: 11477. DOI: 10.1038/s41598-025-96508-3

Bojić, Ljubiša, Predrag Kovačević, and Milan Čabarkapa. 2025. "Does GPT-4 Surpass Human Performance in Linguistic Pragmatics?" *Humanities and Social Sciences Communications* 12: 794. DOI: 10.1057/s41599-025-04912-x

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Braun, Virginia, and Victoria Clarke. 2006. "Using thematic analysis in psychology." *Qualitative Research in Psychology* 3 (2): 77–101.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, SJ Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei [Brundage *et al.*]. 2018. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv*1802.07228. DOI: 10.48550/arXiv.1802.07228

Brynjolfsson, Erik, and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W. W. Norton.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Proceedings of Machine Learning Research* 81: 1–15.

Cath, Corinne. 2018. "Governing artificial intelligence: ethical, legal and technical opportunities and challenges." *Philosophical Transactions of the Royal Society* 376 (2133): 20180080.

Chui, Michael, James Manyika, and Mehdi Miremadi. 2016. "Where Machines Could Replace Humans—and Where They Can't (Yet)." *McKinsey Quarterly*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet

De Filippi, Primavera, Samer Hassan, and Roberto Zicari. 2023. "Smart contracts meet legal responsibility: The role of Reg-LLMs in adaptive law." *Stanford Journal of Blockchain Law & Policy* 6 (1): 77–104.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a rigorous science of interpretable machine learning." *arXiv* 1702.08608. DOI: 10.48550/arXiv.1702.08608

Esteva, Andre, Brett Kuprel, Roberto Novoa, Justin Ko, Susan Swetter, Helen Blau, Sebastian Thrun [Esteva *et al.*]. 2017. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542 (7639): 115–118. DOI: 10.1038/nature21056

Feldstein, Steven. 2019. "The Global Expansion of AI Surveillance." Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance?lang=en

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Lütge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, Effy Vayena

[Floridi *et al.*]. 2018. “AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations.” *Minds and Machines* 28 (4): 689–707.

Gabriel, Iason. 2020. “Artificial intelligence, values, and alignment.” *Minds and Machines* 30 (3): 411–437.

Helberger, Natali. 2019. “ On the democratic role of news recommenders.” *Digital Journalism* 7 (8): 993–1012.

Hinton_60Minutes. 2023. “Godfather of AI” Geoffrey Hinton: The 60 Minutes Interview. *60 Minutes. Youtube.* https://youtu.be/qrvK_KuIeJk

Hinton_Diary. 2025. “Godfather of AI: I Tried to Warn Them, But We’ve Already Lost Control! Geoffrey Hinton.” *The Diary of A CEO. Youtube.* https://youtu.be/giT0ytynSqg

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The global landscape of AI ethics guidelines.” *Nature Machine Intelligence* 1 (9): 389–399.

Korinek, Anton, and Joseph E. Stiglitz. 2018. “Artificial intelligence and its implications for income distribution and unemployment.” In *Economics of Artificial Intelligence: An Agenda*, eds. Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 259–290. Chicago: University of Chicago Press.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. “ImageNet classification with deep convolutional neural networks.” *Communications of the ACM* 60 (6): 84–90.

LeCun_Brian. 2024. “Yann LeCun: AI Doomsday Fears Are Overblown [Ep. 473].” *Dr Brian Keating. Youtube.* https://youtu.be/u7e0YUcZYbEhttps://youtu.be/u7e0YUcZYbE

LeCun_Lex. 2024. “Yann Lecun: Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI | Lex Fridman Podcast #416.” *Lex Fridman. Youtube.* https://youtu.be/5t1vTLU7s40https://youtu.be/5t1vTLU7s40

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. “Deep learning.” *Nature* 521 (7553): 436–444.

Metz, Cade. 2023. “‘The Godfather of A.I.’ Leaves Google and Warns of Danger Ahead.” *The New York Times*. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

Moe, Hallvard, Jan Fredrik Hovden, and Kari Karppinen. 2021. “Operationalizing Exposure Diversity.” *European Journal of Communication* 36 (2): 148–67. DOI: 10.1177/0267323120966849

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, Sendhil Mullainathan [Obermeyer *et al.*]. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366 (6464): 447–453.

O'Connor, Cailin, and James Owen Weatherall. 2019. *The Misinformation Age: How False Beliefs Spread*. New Haven: Yale University Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books.

OpenAI. 2023. "GPT-4 Technical Report." *OpenAI*. https://cdn.openai.com/papers/gpt-4.pdf

OSF 2025. "Risk and Responsibility at the Frontier of AI." *OSF*. https://osf.io/s4gva/?view_only=b42e04616959401082178f6c4c8376ce

Pariser, Eli. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. New York: Penguin Press.

Pavlovic, Maja, and Ljubisa Bojic. 2020. "Political Marketing and Strategies of Digital Illusions – Examples from Venezuela and Brazil." *Sociološki pregled* 54 (4):1391–1414. DOI: 10.5937/socpreg54-27846

Reinhardt, Anne, Jörg Matthes, Ljubisa Bojic, Helle T. Maindal, Corina Paraschiv, and Knud Ryom [Reinhardt *et al.*]. 2025. "Help Me, Doctor AI? A Cross-National Experiment on the Effects of Disease Threat and Stigma on AI Health Information-Seeking Intentions." *Computers in Human Behavior* 172: 108718. DOI: 0.1016/j.chb.2025.108718

Rudin, Cynthia. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1 (5): 206–215.

Russell, Stuart J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, Cedric Langbort. [Sandvig *et al.*]. 2016. "Auditing algorithms: Research methods for detecting discrimination on Internet platforms." *Data and Discrimination: Converting Critical Concerns into Productive Inquiry. Social Science Research Council*. https://ai.equineteurope.org/system/files/2022-02/ICA2014-Sandvig.pdf

Schmidhuber, Jürgen. 2015. "Deep learning in neural networks: An overview." *Neural Networks* 61: 85–117.

Silver, David, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, Demis Hassabis [Silver *et al.*]. 2016. "Mastering the game of Go with deep neural networks and tree search." *Nature* 529 (7587): 484–489.

Standing, Guy. 2018. *Basic Income: And How We Can Make It Happen*. London: Pelican Books.

Susskind, Daniel. 2020. *A World Without Work: Technology, Automation, and How We Should Respond*. New York: Metropolitan Books.

Turing. 2018. "Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award. Bengio, Hinton and LeCun Ushered in Major Breakthroughs in Artificial Intelligence." *Association for Computing Machinery*. https://awards.acm.org/about/2018-turing

Turing_Bengio. 2018. "Yoshua Bengio." *AM Turing Award*. https://amturing.acm.org/award_winners/bengio_3406375.cfm

Turing_Hinton. 2018. "Geoffrey Hinton." *AM Turing Award*. https://amturing.acm.org/award_winners/hinton_4791679.cfm

Turing_LeCun. 2018. "Yann LeCuno." *AM Turing Award*. https://amturing.acm.org/award_winners/lecun_6017366.cfm

West, Sarah Myers. 2019. "Data Capitalism: Redefining the Logics of Surveillance and Privacy." *Business & Society* 58 (1): 20–41.

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Љубиша Бојић[*]

Истраживачко-развојни институт за вештачку интелигенцију Србије, Нови Сад

Институт за филозофију и друштвену теорију, Универзитет у Београду

Лабораторија за дигитално друштво, Београд

Complexity Science Hub, Беч, Аустрија

РИЗИК И ОДГОВОРНОСТ НА ГРАНИЦИ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ: ТЕМАТСКА АНАЛИЗА СТАВОВА ПИОНИРА ДУБОКОГ УЧЕЊА О ПРЕТЊАМА И УПРАВЉАЊУ ВИ[**]

Резиме

Ова студија приказује резултате квалитативне тематске анализе интервјуа са три најзначајнија пионира дубоког учења – Џефријем Хинтоном, Јошуом Бенџијом и Јаном ЛеКуном, добитницима Тјурингове награде, у вези са ризицима вештачке интелигенције (ВИ) и изазовима управљања. Анализа издваја седам главних тематских кластера: егзистенцијалне и дугорочне опасности, друштвене ризике (запосленост, неједнакост), злоупотребе (кибер, биолошко оружје), друштвено-психолошке утицаје (манипулација, поларизација), питања управљања и регулације, позитивне потенцијале, те лична осећања и одговорности научника. Хинтон и Бенџио наглашавају егзистенцијалне ризике и потребу за глобалном регулативом, недвосмислено упозоравајући на потенцијал изумирања људске врсте у уколико не буде адекватних механизама за контролу и усклађивање вредности ВИ са људским интересима. Бенџио додатно показује еволуцију свог става, прелазећи од техно-оптимизма ка осећају друштвене дужности

[*] Имејл: ljubisa.bojic@ivi.ac.rs; ORCID: 0000-0002-5371-7975

[**] Ова студија реализована је уз подршку Министарства науке, технолошког развоја и иновација Републике Србије према Уговору о реализацији и финансирању научноистраживачког рада број 451-03-136/2025-03/ 200025 и Уговору о реализацији и финансирању научноистраживачког рада број 451-03-136/2025-03

и опреза. ЛеКун, са друге стране, одбацује катастрофичне сценарије и заговара децентрализован, отворен развој, верујући у техничке и друштвене капацитете да се ризици успешно савладају, манифестујући снажан оптимизам. Сви саговорници препознају ризик од социјалних подела кроз манипулативне алгоритме, као и потенцијал масовне незапослености и концентрације богатства, али се разликују у хитности и врсти препоручених решења. Хинтон и Бенџио сугеришу регулисање на нивоу инфраструктуре и увођење нових глобалних институција, док ЛеКун заступа постепене техничке и друштвене адаптације. Међу иновативним предлозима за управљање ВИ издвајају се "*compute* пасоши", условне лиценце за напредне моделе, професионалне дозволе за истраживаче и увођење разноврсности садржаја на великим платформама за борбу против негативних алгоритамских утицаја. Посебно је наглашена потреба за адаптивном регулативом која би могла да прати развој технологије, као и за јачањем професионалне одговорности и друштвене свести. Резултати указују да експертска мишљења о ВИ ризику нису једногласна, већ прожета значајним дискусијама о степену претње и приоритетима у политици и регулисању ВИ. Закључује се да су иновативне, интернационално координиране регулаторне интервенције неопходне ради балансирања потенцијала и претњи ВИ, те да професионална и друштвена одговорност развојних инжењера мора постати кључна компонента новог модела управљања. Будућа истраживања треба да обухвате шири спектар актера, укључујући различите друштвене, политичке и међународне перспективе, и тестирање предложених механизама у реалним условима, имајући у виду да се природа ризика динамично мења са ширењем и усвајањем ВИ у новим доменима.

Кључне речи: вештачка интелигенција, пионири дубоког учења, закони који регулишу ВИ, егзистенцијалне претње, тематска анализа

[*] This manuscript was submitted on July 5, 2025, and accepted by the Editorial Board for publishing on November 10, 2025.