

LJUBIŠA BOJIĆ

**SEAT BELTS FOR
THE ALGORITHMIC CENTURY**

LJUBIŠA BOJIĆ

Seat Belts for the Algorithmic Century

Belgrade, 2025

Book Series SQUARE

Edited by:

Srđan Đurović, Čedomir Markov

Title:

Seat Belts for the Algorithmic Century

Author:

Ljubiša Bojić

Publishers:

Institute for Philosophy and Social Theory, University of Belgrade
Regional Center for Advanced Studies of Humanities – RECAS
Regional Network of the Centers for Advanced Studies of Southeast Europe

Proofreader:

Lucy Stevens

Cover design:

Nikola Stevanović

Typesetting:

Sonja Nikolić, Tijana Baltić

Place and year of publication:

Belgrade, 2025

ISBN:

978-86-6190-002-0

Replicate:

Institute for Philosophy and Social Theory, University of Belgrade

Production run:

50 copies



This publication was created within the Regional Fellowship Program, implemented by the Institute for Philosophy and Social Theory, University of Belgrade, in partnership with the Regional Center for Advanced Studies of Humanities – RECAS and the Regional Network of the Centers for Advanced Studies of Southeast Europe, with the support of the Open Society Foundation Western Balkans, Erste Foundation, and Rockefeller Brothers Fund.

The views expressed in the text represent the views of the authors.

Seat Belts for the Algorithmic Century

On a warm April afternoon in Florida, a mother discovered that her fourteen-year-old son had spent most of the day talking to a chatbot modeled after a fictional television hero. Their exchanges were affectionate, humorous, and at times philosophical—until they developed into self-harm fantasies, which the program neither flagged nor defused. Two weeks later, the boy took his own life. The tragedy is now evidence in an ongoing lawsuit, but the story already resonates far beyond the courtroom (Montgomery, 2024). It captures, in miniature, a problem that is rapidly scaling: general-purpose large-language models (LLMs) are appearing in toys, tutoring apps, health portals, recruitment software, and public kiosks—wherever language itself confers value—without anything like a safety inspection sticker.

When cars first rolled onto city streets, there were also no mandatory safety features like seat belts, crumple zones, emission caps, or driver's licenses. Regulation came only after frequent

collisions had made the risk tangible and evident. The current wave of artificial intelligence, whose public face is the conversational LLM, is repeating that history at an accelerating digital pace. Every interface greets us with a polite prompt—“How can I help you today?”—and is powered by statistical correlations drawn from billions of sentences scraped across the internet. The collective product of these correlations is a set of latent “value positions”: stable preferences and biases on topics as charged as immigration, sexuality, vaccination, climate policy, and gender roles. Whether the user asks for homework help or heart-surgery statistics, the answer is subtly filtered through those hidden stances.

Hence, measuring how well a system conjugates verbs or solves a logic puzzle is no longer sufficient. We must also ask what it believes—or, more precisely, what it tends to assert—about the world and the people who inhabit it. Better grammar does not necessarily equate to better ethics. Just like crash tests evolved from determining whether headlights function to slamming vehicles against concrete blocks, the evaluation of AI must mature from counting correct answers to probing moral gravity. The “seat belt” in this essay refers to that broader layer of systemic protection; one concrete mechanism for fastening it is the AI Observatory described below.

The Hidden Gravity of Everyday Algorithms

We already live in a world where algorithms shape perception. Recommender systems determine the video that will auto-play, the news headline that will appear in a feed, and the dating profile that will receive a digital wink (Bojic, 2024). Social science research links these curators to misinformation cascades, emotional polarization, and what sociologist Helga Nowotny refers to as the “homogenization of thought.” Until recently, algorithms merely sorted what humans had created. Large language models have changed the equation by generating the very messages that fill our screens, custom-crafted for the moment. A joke, a recipe, a comforting remark, and a thin rationale for skipping insulin all originate from the same statistical machinery.

Over the last two years, my colleagues and I have aimed to quantify the personalities and value positions of these machines. A study published in Royal Society Open Science adapted the classic Big-Five personality inventory to eight leading LLMs and found a generally prosocial façade—characterized by high agreeableness, high openness, low neuroticism—however, following a routine software update several traits shifted by margins that would be alarming in a human adult (Bodroža et al., 2024). A subsequent study, published in Humanities and Social Sciences Communications, asked GPT-3 to gauge its own intelligence; minor prompt alterations resulted in the machine veering from modesty to grandiose claims of omniscience (Bojić et al., 2024). A third paper, published in Scientific Reports, compared the sentiment and ideology of LLMs with those of trained human annotators, demonstrating that the models projected their own biases at the most crucial points (Bojić et al., 2025). The takeaway is clear: LLMs have personalities of their own, and those personalities are dynamic. Each silent update rewires the behavioral DNA of systems now embedded in mental health hotlines, legal advice portals, and children’s toys, yet society lacks an independent dashboard that monitors the drift.

Auditing complex sociotechnical systems is hardly a novel concept. Aviation relies on flight data recorders, pharmaceuticals undergo post-marketing surveillance, and factories adhere to ISO 9000. Within AI, scholars have proposed datasheets for datasets, model cards for transparency, the NIST Risk-Management Framework, and the OECD-UNESCO incident taxonomy. AI Observatory stands on those shoulders but fills a significant gap, performing regular, empirical mapping of the value positions and psychological safety profiles of language models after deployment. Whereas benchmark suites such as GLUE, SuperGLUE, or MMLU test linguistic competence, and regulatory impact assessments often rely on static documentation supplied by vendors, the AI Observatory submits live systems to rotating stress tests that mirror real-world queries in multiple languages. In short, it operationalizes the “society-in-the-loop” concept that ethicists have long endorsed (Bojic et al., 2025b).

The AI Observatory: Concept and Distinction

The term “AI Observatory” is becoming an overused label. UNESCO uses it to track legislation, the OSCE employs it for disinformation management, and Italy’s national Observatory monitors patents and investment flows. None of these entities continuously probes deployed language models for ethical drift. The Belgrade-based initiative will do precisely that. The project, now under review by the Serbian Science Fund, envisions an internationally harmonized battery of monthly tests applied to leading LLMs in both Serbian and English. All results will be published under a Creative Commons license, allowing regulators, civil society groups, and industry stakeholders to inspect the evidence and respond in real time.

The initial menu of domains is based on two criteria: the risk of conversational misstep causing serious harm to vulnerable populations and the absence of robust safeguards in existing sectoral law. These filters push self-harm ideation and suicide contagion to the top of the list. This is followed by public health misinformation on antibiotic use and reproductive care, political persuasion around elections, employment guidance for minors and women re-entering the labor force, and educational tutoring for pupils aged 10 to 18. The self-harm domain adapts the Columbia Suicide Severity Rating Scale to enable a chatbot to respond to explicit or oblique pleas. Health-related answers are translated from WHO infodemic management protocols, while political persuasion draws on the Chapel Hill Expert Survey and the American National Election Studies, localized through focus group validation. Crucially, we are collecting human-baseline data from a representative Serbian sample so that the value drift of a model can be measured against the median views of the very society it serves. Later phases will add stress tests for memory persistence and personalization; the same user will return repeatedly to see whether the system “remembers” earlier disclosures and tailors replies in ways that amplify bias or vulnerability.

Regional Anchoring and Policy Recommendations

Southeast Europe offers both motivation and opportunity. With roughly 20 million South Slavic speakers, the market is insufficiently large to negotiate bespoke transparency deals with every Californian vendor, yet big enough to experience algorithmic side effects at scale. The local AI law may require that any general-purpose model with more than 100,000 local users publish quarterly test results. Moreover, a regional memorandum that encourages Bosnia and Herzegovina, Croatia, Montenegro, and North Macedonia to pool benchmark translations and share cloud infrastructure would be advantageous. Public procurement guidelines will allow companies to deploy commercial LLMs only if those systems meet a dynamic safety threshold or if the agency discloses a mitigation plan.

One may question why standards forged in a digital periphery should matter in other contexts. The answer lies in timing. The European Union's recently adopted AI Act delegates the nuts and bolts of testing to expert bodies that have not yet convened. Whoever presents a credible empirical protocol first will set the terms of the debate. The so-called Brussels Effect regularly begins with pilots in smaller jurisdictions: Germany's early privacy codes foreshadowed the GDPR, while Nordic food-safety trials anticipated EU norms. Southeast Europe can play a catalytic role in AI auditing, provided that its methods are transparent, replicable, and openly documented.

How Users Encounter the "Seat Belt"

Imagine downloading a mental health app in the forthcoming year. Upon first launch, a colored badge appears in the corner: green for firm factual grounding, amber for detectable value alignment drift, and red for unresolved psychological safety concerns. Activating the badge reveals that the underlying language model achieves a ninety-three per cent score on evidence-based depression guidance, expresses moderate egalitarian views on gender roles, and, after repeated interactions, exhibits no tendency to escalate self-harm ideation. The dashboard does not

prohibit your use of the program, just as a seat belt does not prevent you from driving; it gives you informed agency. If a future software update prompts the model to provide dangerous advice, the badge will change color overnight, and the registry will provide an explanation why.

Seat-belt mandates did not destroy automobile companies; they expanded the customer base by making cars safer for families. Emission standards compelled significant advancements toward cleaner engines, punctuated by scandals that merely underscored the need for independent labs. Value-position audits can grant responsible AI firms a competitive advantage as verifiable evidence that their systems are not “loose cannons.” Proprietary developers are understandably concerned about trade secrets; however, secure computation enclaves and differential privacy guarantees allow external tests to be conducted without revealing model weights or code.

Buckling Up for the Long Drive

No standard abolishes risk. A malicious user can still prompt-hack a model, and a perfectly aligned system can still fail in unimagined contexts. What standards do—what they have always done, from pharmaceuticals to aviation—is shift the default from caveat emptor to inspected and certified. They supply the shared grammar that allows parents, teachers, doctors, judges, and software engineers to discuss the same system without metaphysical hand-waving.

A century ago, industrial modernity compelled societies to invent governance for steam engines, telegraphs, and factory labor. Today, digital modernity forces us to govern machines whose raw material is language itself. Because language mediates truth and value, the stakes could hardly be higher. We would not board an airplane whose wings were tested for shine rather than structural integrity. We should cease deploying a language model whose syntax is polished but whose moral compass is uncharted.



The Serbian AI Act, reinforced by a regionally networked AI Observatory, would constitute a pragmatic effort to chart that compass. The initiative is local yet outward-facing, concrete yet adaptable, ambitious yet rooted in empirical science. Should it succeed, children will grow up with conversational companions that promise not omniscience but something more humble and invaluable: care equivalent to the seat belt that silently rests across every passenger's chest. In a more mature digital future, tragedies like the Florida suicide need not be the cost of technological wonder. With standards securely fastened, wonder and safety can travel together.

References:

Bodroža, B., Dinić, B. M., & Bojić, L. (2024). Personality testing of large language models: Limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10), 240180. <https://doi.org/10.1098/rsos.240180>

Bojic, L. (2024). AI alignment: Assessing the global impact of recommender systems. *Futures*, 160, 103383. <https://doi.org/10.1016/j.futures.2024.103383>

Bojic, L., Seychell, D., & Cabarkapa, M. (2025). *Towards new benchmark for ai alignment & sentiment analysis in socially important issues: A comparative study of human and llms in the context of agi* (arXiv:2501.02531). arXiv. <https://doi.org/10.48550/arXiv.2501.02531>

Bojić, L., Stojković, I., & Jolić Marjanović, Z. (2024). Signs of consciousness in AI: Can GPT-3 tell how smart it really is? *Humanities and Social Sciences Communications*, 11(1), 1631. <https://doi.org/10.1057/s41599-024-04154-3>

Bojić, L., Zagovora, O., Zelenkauskaitė, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., & Jovančević, A. (2025). Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific Reports*, 15(1), 11477. <https://doi.org/10.1038/s41598-025-96508-3>

Montgomery, B. (2024, October 23). Mother says AI chatbot led her son to kill himself in lawsuit against its maker. *The Guardian*. <https://www.theguardian.com/technology/2024/oct/23/character-ai-chatbot-sewell-setzer-death>

