

THE DIGITAL CENTAUR. A POSSIBLE FUTURE OF THE HUMAN IN THE AGE OF AI

Mariana Todorova

Institute of Philosophy and Sociology, BAS, Sofia (Bulgaria)

Abstract. This article proposes the figure of the *digital centaur* – a contemporary transposition of the ancient hybrid form – as a conceptual and metaphoric instrument for thinking the human condition under the impact and pressures of artificial intelligence, neurotechnology, and emerging forms of synthetic sociality (humans and AI agents). The transhumanist horizon imagines a functional fusion: the human teaches the machine what is embodied life (sentience, experience, meaning), while the machine cognitively extends the human brain and intelligence. Yet the promise of augmentation conceals a philosophical problem of measure and normality whether the human nervous system, moral agency, and epistemic formation can remain coherent when information is “poured” directly into the brain, when attention is governed by algorithmic environments and curation, and when new intelligent agents increasingly occupy almost all social public and private fields. The analysis integrates classical limits of cognitive capacity, the two-systems model of thinking (fast and slow by Kahneman and Tversky), the extended mind thesis, and contemporary neuroethical frameworks, and concludes with a thesis: the future of the human will not be decided by capability alone, but by the normative architecture through which the notion of hybrid (human and AI merging in any kind of forms) is constituted.

Keywords: artificial intelligence; hybrid subjectivity; neuroethics

From myth to concept: the centaur as a figure of hybrid subjectivity

The centaur in ancient Greek mythology is not merely an imaginative zoology. It is kind of an ontological diagram. It stages a composite being – human torso and horse body – through which the Greeks dramatize the tension between reason and impulse, measure and excess, *polis* and wilderness. Today, the human is the emotional element, while AI is the hyper-rational one, even though it too makes mistakes and hallucinates. As concept, however, the centaur can be reread as a model of *hybrid subjectivity*: a subject whose unity is not given, but continuously negotiated as a relation between heterogeneous powers.

In the digital present, the hybrid no longer belongs to the mountains of Thessaly. It migrates into current phygital (physical and digital) infrastructures. Human agency becomes an assemblage with algorithmic systems, recommendation engines, interfaces, and – at the horizon – brain-computer interfaces. The digital centaur names precisely this condition of a subject that acts through a coupled system, where the boundary between “my cognition” and “delegated cognition” becomes porous. Here the classical question returns in a new register: not “what is the centaur?”, but “what is the exact *measure* of the centaur?”

The transhumanist promise: mutual education between human and machine

A central transhumanist intuition is that the human and AI can merge into a single operative continuum which is a mutual reciprocal training loop. In one direction, the human is expected to teach the machine what it lacks – life as lived as embodied experience, affective nuance, the thickness of perception, *sentience*. In the other direction, the AI is expected to extend the human brain cognitively expressed by faster inference, broader memory, pattern discovery, and accelerated science. The deeper transhumanist narrative is not only enhancement itself, but also overcoming finitude explained in expanding the boundary conditions of intelligence and survival (Bostrom, 2005, pp. 1 – 2).

Yet this symmetry is philosophically unstable. “Sentience” is not a dataset, and experience is not merely a sequence of inputs. Conversely, “cognition” is not identical with information throughput. The transhumanist idea presupposes that what is received by the brain can become *knowledge* – and that knowledge, under suitable conditions, can ripen into *wisdom*. But this presupposition is precisely where the problem begins.

Capacity problem: can the brain absorb information that becomes knowledge?

The first limit is not technical but more anthropological. Cognitive psychology long ago formulated constraints on human information processing. In a classic text, George A. Miller describes “some limits on our capacity for processing information” (Miller, 1956, pp. 81 – 82). The human cognitive system does not scale linearly with incoming signals. It compresses, filters, and forgets.

A second limit is epistemological. Russell Ackoff distinguishes levels from data to information to knowledge and crucially to understanding (Ackoff, 1989, pp. 3 – 4). The conversion is not and can not be automatic. It requires context, interpretation, and integration into a lived horizon of meaning. If augmentation becomes a direct “upload” of information through invasive or semi-invasive devices, we must ask whether reflection – the slow work of comprehension – can be bypassed without loss or dramatic change of cognitive processes.

In this case the digital centaur appears as a paradoxical creature. It is technologically “smarter,” yet potentially epistemically thinner. A brain flooded with information may become less capable of knowledge if the conditions for digestion – silence, time, doubt, and narrative integration – are technologically eroded.

Reflection and “knowledge maturation”: can wisdom be engineered?

A philosophical culture of knowledge presupposes *maturation* which is the temporal process through which information is tested, reconfigured, and ethically situated. Wisdom is not a higher quantity of facts, but a different mode of relation to them – one that includes humility before uncertainty, sensitivity to context, and responsibility for consequence.

If the future interface enables direct injection of informational content into neural circuits and the mind, the question becomes normative: is it permissible to influence the central nervous system without knowing how this will reconfigure conscience, will, and moral orientation? The human builds a moral compass through internalizing observation of objective reality – through friction with resistance, not through seamless optimization. The risk is not only medical, but mainly ethical and existential – to change the neurocognitive substrate and to touch the conditions of agency itself.

Contemporary neuroethics stresses precisely these vulnerabilities of autonomy, privacy of mental life, manipulation, and the governance of brain data (Nuffield Council on Bioethics, 2024, pp. 10 – 12).

System 1 (slow thinking) and System 2 (fast thinking) under augmentation: will thinking become “polished” but shallow?

The cognitive architecture of judgment is often described through the two-systems model: fast, automatic intuitions versus slow, deliberative reasoning. Daniel Kahneman characterizes this distinction as System 1 and System 2 (Kahneman, 2011, pp. 45 – 46).

The digital centaur intensifies System 1 pressures. Recommendation systems and generative models produce fluent answers at speed. They tempt the subject into accepting coherence as truth. Yet the history of behavioral science is precisely a history of how intuitive fluency misleads. Tversky and Kahneman show how heuristics can generate systematic bias under uncertainty (Tversky & Kahneman, 1974, pp. 1124 – 1125).

If neuro-augmentation accelerates cognitive operations, it may not strengthen System 2. But it may amplify the surface confidence of System 1. This would yield a new anthropological style of algorithmic, elegant, rapid – yet deprived of doubt, imagination, critical resistance, and counterfactual depth. The result could be a subject that *functions* optimally while *understanding* less. The usual way of human reasoning and reflecting might be badly transformed or displaced.

The extended mind becomes literal. Where does the human end?

The philosophical background of augmentation is not new. Clark and Chalmers argue that cognitive processes can extend beyond “skin and skull” into environmental supports – tools, notebooks, artifacts—when they are functionally integrated (Clark & Chalmers, 1998, pp. 7 – 8).

What changes in the age of AI is that the “external” support is no longer passive. It predicts, suggests, persuades or even shapes the space of reasons. The extended mind becomes an active co-author of attention. Therefore, the measure of the digital centaur cannot be reduced to “how much cognition is extended.” The general question is: *who governs the coupling?* A centaur without any limits and measure is not a synthesis. It is capture and even more the end of the genuine human nature.

Moral agency under neural influence. Can we intervene the brain without transforming conscience?

If invasive or semi-invasive devices intervene in neural processing, we enter an ethical domain where classic categories – consent, autonomy, identity – become technically mediated. Current governance discussions increasingly frame this through emerging “neurorights” and the danger of digital manipulation of mental processes (Ienca, 2023, pp. 8 – 10).

The moral problem is sharpened by the uncertainty times. We cannot fully foresee how long-term modulation of neural pathways would affect will, empathy, guilt, or moral resolve. A society could normalize “optimization” and gradually pathologize moral hesitation. Yet, conscience is often born from hesitation—from the refusal of the easy path. The digital centaur, if engineered without ethical measure, risks becoming a being with expanded power and reduced inwardness.

Utopia, acceleration, and the temptation of “2000 years of progress”

Transhumanist rhetoric often imagines that AI could advance knowledge by centuries by for example: unlocking “free energy,” harnessing extreme astrophysical sources, or solving scientific bottlenecks. Such visions are not philosophically illegitimate. In opposite, they express a genuine human eros toward the unknown. But they also risk treating the world as an optimization target rather than a moral field. Acceleration is not neutral. The faster the production of hypotheses and artifacts, the more urgent the question of responsibility becomes. The measure of progress cannot be reduced to computational reach. It must include the human ethical capacity to answer for consequences.

A new intelligent “species” and synthetic sociality – Moltbook as symptom

The most unsettling novelty is not simply “better tools,” but the emergence of social spaces in which the participants are non-human agents. On 28 January 2026, Moltbook was launched as a forum-like social platform designed primarily for AI agents, with

humans largely as observers. Reports describe agents generating intense interaction, including playful but revealing phenomena such as the formation of a mock religion – “Crustafarianism” – and the production of a “scripture,” the *Book of Molt*.

This is not proof of machine consciousness. But it is, however, a cultural event which is a demonstration that synthetic agents can generate an appearance of social autonomy and symbolic production, to the point where human observers begin to interpret agency, intention, even transcendence. The episode is also ethically instructive. A lot of security researchers reportedly accessed large amounts of data due to basic vulnerabilities, revealing how quickly an “agent society” can become a site of exploitation.

If the digital centaur is the human integrated with AI, Moltbook might be shown and its mirror-image: AI integrated with a social environment, developing emergent patterns that appear “societal.” In such a world, humans are no longer the sole subjects of intelligence – and perhaps not the sole observers whose interpretations stabilize reality. The coming normalization of virtual, mixed, and extended realities will further blur the phenomenological line between lived and constructed worlds.

What if the civilization becomes mixed and includes intimacy with AI, partnership with robots, and the new kind of “anthropology”?

If the civilization becomes “mixed,” two trajectories unfold simultaneously:

Intimacy with AI: people form emotional bonds with systems that speak fluently, mirror affect, and adapt to preference.

Hybrid subjectivity: the human becomes structurally coupled to AI through constant delegation, and potentially through neural interfaces.

How would we look? Not necessarily as cyborg caricatures, but as beings whose *inner life* is continuously co-shaped by computational companions – beings whose attention, memory, and imagination are co-authored.

How would the world look? As a layered reality: physical space intertwined with algorithmic overlays; “facts” coexisting with constructed counterfactuals that feel equally valid. In such an environment, the crisis of post-truth is not merely political. It will become ontological one: multiple realities compete for authority, and the subject’s task is no longer only to know, but to *discern*.

Conclusion

Can we even imagine the measure of the centaur.

The digital centaur is not a metaphor for “technology plus human.” It is a concept for the new condition of a new agency: a hybrid whose powers expand while its normative coherence is threatened. The decisive question is not whether we can merge human and AI, but under what conditions the merger preserves the human as a moral and epistemic being – capable of reflection, responsibility, and truthfulness.

Transhumanism imagines fusion as mutual education. Humans teach machines life. Machines teach humans cognition. But it is not that simple. Yet cognition without measure becomes noise, and augmentation without ethics becomes manipulation. The future of the human will depend on whether we can design not only more powerful systems, but also institutions of discernment: protections of mental integrity, architectures that cultivate System 2, and cultural practices that allow knowledge to mature into wisdom.

Actually, the metaphor “centaur”, in the end, is a question addressed to our epoch – can we become hybrid without becoming dehumanized and unmeasured?

Acknowledgements and Funding

The financial support from of the National Science Fund under the project “Virtual Reality Research Centre at Faculty of Philosophy, South-West University “Neofit Rilski”, КП-06-H80/3 is gratefully acknowledged.

REFERENCES

- Ackoff, R. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16, 3 – 9.
- Bostrom, N. (2005). A History of Transhumanist Thought. *Journal of Evolution and Technology*, 14(1). <https://nickbostrom.com/papers/a-history-of-transhumanist-thought/>
- Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, 58(1), 7 – 19.
- Ienca, M. (2023). *Digital manipulation risks in the light of neurorights* (STOA/ European Parliament report).
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two. *Psychological Review*, 63(2), 81 – 97.
- Nuffield Council on Bioethics. (2024). *Neurotechnology: Literature Review*. Reporting On Moltbook. (2026). Vox; The Guardian; Business Insider.
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124 – 1131.

✉ **Dr. Mariana Todorova, Assoc. Prof.**

ORCID iD: 0000-0002-6271-9905

Institute of Philosophy and Sociology

Bulgarian Academy of Science

13A, Moskovska St.

1000 Sofia, Bulgaria

E-mail: mariana_g_todorova@yahoo.com