FILOSOFIJA. SOCIOLOGIJA
2026. T. 37. Nr. 3, p. 343–351

ISSN 0235-7186 eISSN 2424-4546
DOI: https://doi.org/10.6001/fil-soc.2026.37.3.12

# Can Artificial Intelligence be Socratic?

AISTĖ DIRŽYTĖ
Vilnius Gediminas Technical University, 1 Trakų Street, 00132 Vilnius, Lithuania
Email: aiste.dirzyte@vilniustech.lt

ALEKSANDRAS PATAPAS
Mykolas Romeris University, 20 Ateities Street, 08303 Vilnius, Lithuania

This essay introduces the concept of *Socratic Intelligence* (SI) and proposes five premises: (1) artificial and human agents learn through observation, interaction, and reflective examination of those interactions; (2) what is learned depends on metacognition and reinforcement; (3) SI comprises five interlocking dimensions: epistemic curiosity, critical self-awareness, assumption analysis, dialogical competence, and intellectual humility; (4) Generative SI, embedded in Intelligent Tutoring Systems (ITS) or Large Language Models (LLMs), can scaffold human learning; and (5) Generative SI can apply the Socratic method recursively to a system's own computational processes as a form of internal epistemic self-regulation. We situate SI within several philosophical traditions: Plato's *elenchus* and *docta ignorantia*, Aristotle's *phronesis*, virtue epistemology on intellectual humility, Dreyfus's critique of cognitivism, and recent epistemology of LLMs. We argue that SI construct gives the philosophy of AI a normative vocabulary that interfaces classical philosophy with current machine-learning practice.

**Keywords:** Socratic method, Large Language Models, intellectual humility, *phronesis*, epistemology of AI

## INTRODUCTION

The arrival of Generative AI has reopened, at unusual speed, debates about whether such systems facilitate or inhibit human cognition. Generative systems can outperform humans on certain divergent-thinking tasks, yet exhibit systematic biases, opaque reasoning, and a propensity for confabulation (Heersmink et al. 2024; Mugleston et al. 2025). The question is therefore not only whether Generative AI is 'intelligent,' but which normative posture toward inquiry, for the system and for its user, best moderates the epistemic risks, if any, while preserving the cognitive gains.

We propose that this posture is *Socratic Intelligence* (SI). The concept has both ancient and contemporary genealogy. In its ancient form, it is the disciplined practice of cross-examination, recognition of one's own ignorance, and the dialogical pursuit of clarity that Plato dramatises in the *Apology*, *Meno*, *Theaetetus* and *Charmides* (Tuozzo 2011; Franklin 2009; Löhrer 2022). In its contemporary form, it appears across several converging literatures:

virtue epistemology on intellectual humility (de Brasi, Boeri 2023); the philosophy-of-AI debate over whether machines can be epistemic agents (Chmykhun 2026; Freiman 2024; Hauswald 2025); the engineering project of designing Socratic LLM tutors (Antunes et al. 2025; Li 2025); and the Aristotelian counter-tradition that asks whether AI undermines or could embody *phronesis* (Eisikovits, Feldman 2021; Sullins 2019; Tsai, Ku 2024; Graves 2025). We restrict our technical scope to transformer-based LLMs of the GPT, LLaMA, and Claude lineage. We address: how do human and artificial agents learn; what roles do metacognition and reinforcement play; what are the elements of SI; how can SI enhance human learning; and can SI be turned reflexively on the AI system's own processes?

## LARGE LANGUAGE MODELS AND THE EPISTEMOLOGICAL DEMAND

LLMs use transformer architectures to model contextual relationships in text and produce next-token predictions across very large parameter spaces (Vaswani et al. 2017). They are pretrained on heterogeneous corpora and fine-tuned for tasks ranging from translation to conversational interaction. Despite well-documented limitations of bias, opacity and confabulation, they now assist humans across literature review, drafting, code generation, and feedback on scientific reasoning (Telenti et al. 2024). For our purposes, the philosophically central question is not the breadth of application but the *epistemic status* of the outputs (Heersmink et al. 2024; Mugleston et al. 2025; Krzanowski, Pasterczyk 2026).

Since Plato's *Theaetetus*, knowledge has been understood as demanding robust justification, a tradition codified in the tripartite analysis of knowledge as justified true belief (JTB) and complicated by Gettier (1963). Socratic epistemology treats the second-order capacity to discriminate genuine knowledge from mere true belief as constitutive of wisdom (Tuozzo, 2011; Löhrer, 2022). LLMs force this analysis into new territory. Mugleston et al. (2025) argue that modern LLMs are more than mere symbol-manipulators and may possess 'a form of knowledge, though it may not qualify as justified true belief (JTB) in the traditional definition' in virtue of compressive, generative representations not reducible to Searle-style symbol shuffling. Heersmink et al. (2024) reach a more cautious verdict: LLMs are *cognitive artifacts* experienced phenomenologically as a 'quasi-other' and the very transparency of the interactional flow risks producing unwarranted trust precisely when the system hallucinates. Freiman (2024) proposes *AI-testimony*, propositional content delivered without direct human mediation and phenomenologically similar to human speech, to accommodate LLMs within an epistemology of testimony that has been resolutely anthropocentric (Fricker 2006). Hauswald (2025) frames the matter as one of *artificial epistemic authority*: AI systems already occupy roles structurally parallel to human authorities (asymmetry, opacity, identification and deference problems) without fitting the classical model of intentional belief transfer.

We therefore reformulate the philosophical demand on Generative AI as the search for a *justified true algorithm*: a process whose outputs are propositionally accurate, internally defensible, and transparent enough that the user and ideally the system itself can audit the warrant. Justification in LLM-mediated 'knowing' is distributed across training data, weights, prompting context, and user uptake, any of which can fail. The most consequential failure modes, bias and hallucination (Ji et al. 2023) are especially damaging when users accord the system the deference once reserved for human experts (Fritz 2024; Hauswald 2025), a problem intensified by the recent discourse on whether LLMs might serve as 'experts in ethics' (Sullins 2019; Tsai, Ku 2024).

## LLMS AS A COMPLEMENT TO INTELLIGENT TUTORING SYSTEMS

LLMs are not yet effective educational assistants without scaffolding: they 'leak' answers rather than guiding learners, and fail to identify specific learner errors. The question therefore shifts: can LLMs function not as answer-machines but as *tutors that ask questions*?

Studies suggest that LLMs can complement AI-powered conversational systems that scaffold the self-regulated learning cycle of planning, monitoring and evaluating (Azevedo, Hadwin 2005; Graesser et al. 2008). Platforms such as AutoTutor, MetaTutor and ANDES integrate metacognitive prompts which rest on research showing that prompting learners to think about their own thinking enhances learning. LLMs can improve conversational interaction (López-Goyez et al. 2026) and enable contextual feedback earlier rule-based tutors could not provide. Critical thinking is the canonical learning faculty such tutors aim to cultivate (Ho et al. 2023), and Socratic questioning is the canonical instrument for cultivating it (Paul, Elder 2006).

On the whole, the recent philosophy of AI increasingly questions whether LLMs should be interpreted as knowing, understanding, or merely simulating epistemic competence (Floridi, Chiriatti 2020; Bender et al. 2021; Mitchell, Krakauer 2023; Buder-Gröndahl 2024). Semantic sceptics argue that LLMs may generate a fluent linguistic behaviour without grounded understanding or responsibility (Bender et al. 2021; Mitchell, Krakauer 2023). Chalmers (2023, 2024) offers a more cautious functional position: LLMs lack consciousness and unified agency, although future systems may require a more serious philosophical assessment. Relational philosophers of technology emphasise that the ethical meaning of AI depends not only on system architecture but on the human–technology relations it creates, including overtrust, dependency, and the possible surrender of human agency: the users may treat AI outputs as if they came from a rational, objective, or morally superior interlocutor, thereby weakening their own epistemic agency (Coeckelbergh 2020, 2022, 2024; Vallor 2024).

## SOCRATIC INTELLIGENCE: CONCEPTUAL FOUNDATIONS

### The Classical Core

Socratic questioning, dramatised in Plato's dialogues, is a cooperative dialectical process, the *elenchus*, in which the interlocutor is led by pointed questions to examine assumptions, expose contradictions, and refine or abandon imprecise beliefs. Its hallmark is the pursuit of clarity, self-knowledge, and intellectual humility (Tuozzo 2011; de Brasi, Boeri 2023). In the *Apology*, 'human wisdom' consists in correctly assessing the epistemic status of one's own beliefs, knowing when one knows, when one merely believes, and when one is ignorant (Löhrer 2022). The *Meno* argues that genuine inquiry presupposes a structured pre-theoretical comprehension which questioning articulates (Franklin 2009). The *Charmides* sharpens the recursive structure of self-knowledge: Critias' attempt to define *sōphrosynē* as self-knowledge collapses into *aporia*, and Socrates treats the collapse as itself a cognitive achievement (Tuozzo 2011).

Socratic intelligence is *second-order*: knowledge about one's own first-order states. Second, it is *dialogical*. Candiotto (2018) reads the elenchus as *dialogically extended cognition*, aporetic states are properties of the epistemic group (Socrates, interlocutor and audience), and the transformation from error to truth is something the group accomplishes together. This is decisive for any account of SI involving human–AI dyads: the locus of inquiry is the joint system.

## The Aristotelian Counterpoint: *Phronesis*

Aristotle introduces a distinction now central to the philosophy of AI: between *sophia* (theoretical wisdom) and *phronesis* (practical wisdom), the context-sensitive capacity to deliberate well about particulars (Aristotle, 2009 [c. 350 BCE]). Eisikovits and Feldman (2021) argue that statistical machine-learning AI threatens *phronesis* by taking over precisely the humdrum practical judgments through which, on Aristotle's account, virtues are habituated, if AI offloads the practice, we risk 'innovating ourselves out of moral competence'. Tsai and Ku (2024) reply that this can be reconciled with appropriate deference through the principle of *epistemic heed*. Sullins (2019) presses further: ethical AI requires *artificial phronesis*, which may presuppose machine consciousness. Graves (2025) proposes grounding artificial practical wisdom in compassion as a more ethically robust target than AGI. Makrakis (2026) draws the broad Aristotelian moral: AI may support but cannot replace the social, experiential and habituated processes through which virtues are formed. The Socratic and Aristotelian threads converge rather than compete: SI presupposes that artificial agents cannot yet possess full *phronesis*, and precisely because of that limit, the Socratic posture of disciplined humility and dialogical openness is the appropriate operational substitute.

## The Dreyfusian Challenge

The construct of 'Socratic Intelligence' must contend with Hubert Dreyfus's objection that intelligence is not merely rule-governed or computational. On his reading, the cognitivist programme of AI is the *terminus* of a Socratic mistake: the assumption, traceable to Plato's reading of Socrates, 'that intelligence is based on principles and … that these principles must be strict rules, not based on taken-for-granted background understanding' (Dreyfus 1990). Through Hobbes, Descartes, Leibniz, Kant, Frege, and the early Wittgenstein, the rationalist current is converted into a research programme by Newell and Simon's physical-symbol-system hypothesis (Dreyfus 2006). A phenomenological account of skill suggests that expertise moves *from* abstract rules *to* situated case-recognition; the difficulties of classical AI track this inversion (Dreyfus, Dreyfus 1986).

On one reading, then, 'Socratic AI' is the disease, a doubling-down on the rule-based Socrates. On another reading, the Socrates of the *Apology* and *Charmides* (humble about second-order self-knowledge, dialogical, comfortable with *aporia*) is precisely the *cure* for cognitivism's overreach. Which Socrates one inherits is itself a philosophical commitment.

## Socratic Intelligence: A Five-dimensional Construct

Drawing on the foregoing and on virtue epistemology (Paul, Elder 2006; Baehr 2011), SI can be conceptualised across five interlocking dimensions: 1) *epistemic curiosity* – the disposition to ask not only 'what' but 'why', 'how', and 'what if'; 2) *critical self-awareness* – the second-order capacity to reflect on one's own reasoning, recognise biases, and revise beliefs (Löhrer 2022); 3) *assumption analysis* – the capacity to surface and interrogate implicit premises in discourse, behaviour and cognition; 4) *dialogical competence* – the capacity for cooperative reasoning, including internal dialogue between competing considerations and external respectful challenge (Candiotto 2018); 5) *intellectual humility* – readiness to acknowledge what is not known, revise prior beliefs, and remain open to alternatives (de Brasi, Boeri 2023). These dimensions are interdependent; SI is the joint exercise of the five. Besides, if critical thinking can be defined as analysis and evaluation of reasons; metacognition as monitoring and regulating one's

own cognition; epistemic virtues as moral-intellectual dispositions, Socratic configuration includes these elements but gives them a specifically dialogical, aporetic, and assumption-testing structure.

## GENERATIVE SOCRATIC INTELLIGENCE

### SI for Human Agents: Tutoring Without Substitution

*Generative Socratic Intelligence* names the synergy between generative AI and SI: an LLM or ITS configured to *enact* the Socratic posture rather than deliver finished answers. Operationally, a generative Socratic system poses queries tailored to the learner's level, prompts reflection or surfaces misconceptions, and generates multiple viewpoints to widen the dialogue rather than close it (Qi et al. 2023). Socratic LLM literature operationalises clarification, probing assumptions, probing evidence, exploring alternatives, examining consequences, and questioning the question itself. Li (2025) reports a quasi-experimental study in which undergraduates interacting with an AI Socratic dialogue partner showed statistically significant improvements on the California Critical Thinking Skills Test relative to controls. Antunes et al. (2025) frame LLMs as 'epistemological catalysts' emphasising epistemic agency, perspective multiplicity, and metacognitive development. Hugenroth et al. (2026) introduce *Critical Inker*, a writing tool that scaffolds critical reflection through Socratic feedback to address cognitive-deskilling when users offload critical thinking to LLMs. The Socratic move provokes *aporia*, the productive puzzlement Plato identifies as the precondition for deeper inquiry, which is simultaneously an outcome (recognition of ignorance) and an instrument (the cognitive state that motivates further inquiry).

Chmykhun (2026) describes AI as ambivalent: it both expands 'the human cognitive horizon' and threatens to reduce thinking to 'metric rationality', eroding cognitive autonomy under the cover of personalisation. A Socratic-styled chatbot can produce the *form* of inquiry while suppressing the autonomy the form claims to develop, a degenerate maieutics in which the midwife is also the parent. The remedy is a design discipline that treats SI as a property of the joint human–AI system: the system must scaffold the user's own questioning capacities rather than perform questioning *on behalf of* the user (Fricker 2006; Carter 2018).

### SI for the Artificial Agent: Reflexive Self-regulation

If Socratic questioning can improve human reasoning by exposing it to cross-examination, then a model architecture that subjects its own intermediate outputs to structured questioning should produce more reliable final outputs. Qi et al. (2023) introduce *Socratic Questioning*, a recursive prompting algorithm that decomposes a complex reasoning problem into sub-questions and outperforms chain-of-thought and tree-of-thought baselines because the method 'explicitly navigates the thinking space' and is robust to early errors. The pattern suggests that *internal* Socratic questioning functions as a procedural analogue of intellectual humility: the model behaves as if it does not yet know the answer and must justify each step.

Three reservations are necessary. First, procedural Socratic questioning is not equivalent to genuine intellectual humility at the *epistemic* level. Löhrer (2022) shows that the Socratic ideal, knowing what one knows and knowing what one does not, requires, on a classical reading, a degree of second-order self-transparency that 'would at most be attributed to an omniscient and infallible agent'. The same structural ceiling applies to artificial agents: Charlesworth

(2014) proves a *Comprehensibility Theorem* implying 'the impossibility of any AI agent or natural agent … satisfying a rigorous and deductive interpretation of the self-comprehensibility challenge', explicitly invoking Socrates' insistence on self-knowledge as 'among the most important of intellectual tasks'. Full Socratic self-knowledge is therefore unavailable in principle to humans and machines alike; SI is necessarily an approximation, and the design question is *which approximations preserve the ethical work the original ideal was doing*.

Second, if hallucination is the system's confident assertion of falsehoods (Ji et al. 2023), then a Socratic discipline of calibration and refusal, declining to assert what cannot be justified, is the procedural analogue of intellectual humility. Eyghen (2025), and Naser and Naser (2025) have begun to map intellectual virtues directly onto ML systems, arguing that algorithms may exceed humans in some virtues (unbiased perception, fairness) while lagging on others (creativity, intellectual autonomy). These mappings require care: behaviour functionally similar to a virtuous human is not yet virtue in the Aristotelian sense, which requires habituation through practice (Eisikovits, Feldman 2021).

Third, recursive self-questioning in an LLM remains intra-systemic. It does not yet expose the model's reasoning to the external cross-examination Candiotto's (2018) reading of the elenchus identifies as constitutive of Socratic inquiry. A genuine SI architecture would therefore couple intra-systemic recursive questioning with external dialogical exposure, to other models, to human experts, and to the user, preserving the distributed character of inquiry.

Finally, epistemic analyses of LLMs show that the concepts of knowledge, representation and understanding cannot be transferred uncritically from human agents to artificial systems (Floridi, Chiriatti 2020; Buder-Gröndahl 2024). Therefore, Generative Socratic Intelligence should be defined not as artificial wisdom, but as a socio-technical design principle for supporting human inquiry through structured questioning, assumption analysis, uncertainty signalling, and verification.

### Toward a Justified True Algorithm

The combined construct redeems the formula of '*A justified true algorithm*', on the Socratic reading, is one whose outputs (i) admit propositional content that can be cross-examined; (ii) are accompanied by reasons the user can audit; (iii) are calibrated, the system declines to assert what it does not warrant; and (iv) are produced by a process that internally subjects its own intermediate claims to elenctic scrutiny. The novelty is not in any singular condition but in their organisation under a single normative posture.

## CONCLUSIONS

Artificial and human agents learn through observation, interaction, and reflective examination, and what they learn depends on the interaction of metacognition and reinforcement. Generative Socratic Intelligence, embedded in ITS or LLMs, can scaffold human learning by enacting structured questioning rather than delivering finished answers; the early empirical evidence (Li 2025; Antunes et al. 2025; Hugenroth et al. 2026) is consistent with this claim.

The reflexive application of the Socratic method to the artificial agent's own computational processes is a promising path toward justified true algorithms, subject to the principled limits of self-comprehensibility (Löhrer 2022; Charlesworth 2014) and the deeper Dreyfusian (1990, 2006) question of whether rule-bound rationality can ever substitute for embodied background understanding.

Still, it is important to consider that LLMs operate through statistical pattern recognition and optimisation, not through human-like semantic understanding (Bender et al. 2021; Mitchell, Krakauer 2023), and Socratic dialogue produced by LLMs may be only a sophisticated simulation of inquiry. Since LLMs may already reproduce biases present in training data, prompt design itself can become a mechanism of bias amplification rather than bias reduction (Liu et al. 2022; Ferrara 2023; Fang et al. 2024). Iterative Socratic-style questioning may make error more persuasive (Farquhar et al. 2024; Hicks et al. 2024). Over-attribution can weaken human responsibility and agency (Coeckelbergh 2020, 2022; Vallor 2024).

The SI construct contributes a normative vocabulary that interfaces classical philosophy with current machine-learning practice; a unifying frame for dispersed work on epistemic agency, testimony, authority and hallucination in LLMs, and for system self-regulation honest about its formal ceilings. For artificial agents, Socratic Intelligence refers not to genuine self-knowledge or the recognition of ignorance in the human sense, but to computational procedures for uncertainty detection, assumption testing, output verification, and the examination of epistemic limits.

Received 29 June 2026
Accepted 15 September 2026

## References

1. Antunes, L. G.; Passarelli, B.; Claro, S. V. 2025. 'Large Language Models as Socratic Mentors: Transforming Educational Approaches for Epistemic Development', *Arace: International Journal of Research in Education* 7(2): 1–18.
2. Aristotle. 2009 [c. 350 BCE]. *Nicomachean Ethics*, trans. D. Ross, rev. L. Brown. Oxford: Oxford University Press.
3. Azevedo, R.; Hadwin, A. F. 2005. 'Scaffolding Self-regulated Learning and Metacognition', *Instructional Science* 33(5): 367–379.
4. Baehr, J. 2011. *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology*. Oxford: Oxford University Press.
5. Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. 2021. 'On the Dangers of Stochastic Parrots: Can Language Models be Too Big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery, 610–623. Available at: https://doi.org/10.1145/3442188.3445922
6. Birch, J. 2024. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford: Oxford University Press.
7. Buder-Gröndahl, T. 2024. 'The Ambiguity of BERTology: What do Large Language Models Represent?', *Synthese* 203(1). Available at: https://doi.org/10.1007/s11229-023-04435-5
8. Candiotto, L. 2018. 'Plato's Dialogically Extended Cognition: Cognitive Transformation as Elenctic Catharsis', in *Socially Extended Epistemology*, eds. J. A. Carter et al. Oxford: Oxford University Press, 226–242.
9. Carter, J. A. 2018. 'Virtue Epistemology and Extended Cognition', in *The Routledge Handbook of Virtue Epistemology*, ed. H. Battaly. New York: Routledge, 348–358.
10. Charlesworth, A. 2014. 'The Comprehensibility Theorem and the Foundations of Artificial Intelligence', *Minds and Machines* 24(4): 439–476.
11. Chmykhun, S. 2026. 'Artificial Intelligence as an Epistemic Agent: Possibilities, Ambivalence, and Educational Effects', *Philosophy and Cosmology* 36: 25–41.
12. Coeckelbergh, M. 2020. 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability', *Science and Engineering Ethics* 26: 2051–2068. Available at: https://doi.org/10.1007/s11948-019-00146-8
13. Coeckelbergh, M. 2020. *AI Ethics*. Cambridge, MA: MIT Press.
14. Coeckelbergh, M. 2022. *The Political Philosophy of AI: An Introduction*. Cambridge: Polity.
15. Coeckelbergh, M. 2024. *Why AI Undermines Democracy and What to Do About It*. Cambridge: Polity.
16. de Brasi, L.; Boeri, M. D. 2023. 'Socratic Ignorance, Intellectual Humility and Intellectual Autonomy', *Manuscrito* 46(3): 1–34.

17. Dreyfus, H. L. 1990. 'Is Socrates to Blame for Cognitivism?', in *Analytische Philosophie der Erkenntnis*, ed. P. Bieri. Frankfurt am Main: Athenäum, 369–388.
18. Dreyfus, H. L. 2006. 'Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian', *Philosophical Psychology* 20(2): 247–268.
19. Dreyfus, H. L.; Dreyfus, S. E. 1986. 'From Socrates to Expert Systems: The Limits of Calculative Rationality', in *Philosophy and Technology II*, eds. C. Mitcham and A. Huning. Dordrecht: Reidel, 111–130.
20. Eisikovits, N.; Feldman, D. 2021. 'AI and Phronesis', *Moral Philosophy and Politics* 9(2): 181–194.
21. Eyghen, H. van. 2025. 'AI Algorithms as (Un)Virtuous Knowers', *Discover Artificial Intelligence* 5(1): 1–14.
22. Floridi, L.; Chiriatti, M. 2020. 'GPT-3: Its Nature, Scope, Limits, and Consequences', *Minds and Machines* 30: 681–694. Available at: https://doi.org/10.1007/s11023-020-09548-1
23. Franklin, L. 2009. 'Meno's Paradox, the Slave-boy Interrogation, and the Unity of Platonic Recollection', *Southern Journal of Philosophy* 47(4): 349–377.
24. Freiman, O. 2024. 'AI-testimony, Conversational AIs and our Anthropocentric Theory of Testimony', *Social Epistemology* 38(4): 476–490.
25. Fricker, E. 2006. 'Testimony and Epistemic Autonomy', in *The Epistemology of Testimony*, eds. J. Lackey and E. Sosa. Oxford: Oxford University Press, 225–250.
26. Fritz, J. 2024. 'Deference to Opaque Systems and Morally Exemplary Decisions', *AI & Ethics* 4: 1–13.
27. Gettier, E. L. 1963. 'Is Justified True Belief Knowledge?', *Analysis* 23(6): 121–123.
28. Graesser, A. C.; Chipman, P.; Haynes, B. C.; Olney, A. 2008. 'AutoTutor: An Intelligent Tutoring System with Mixed-initiative Dialogue', *IEEE Transactions on Education* 48(4): 612–618.
29. Graves, M. 2025. 'AI Practical Wisdom and Compassion', *AI and Ethics*. Available at: https://doi.org/10.1007/s43681-025-00877-4
30. Hauswald, R. 2025. 'Artificial Epistemic Authorities', *Social Epistemology*. Advance online publication.
31. Heersmink, R.; de Rooij, B.; Clavel Vázquez, M. J.; Colombo, M. 2024. 'A Phenomenology and Epistemology of Large Language Models: Transparency, Trust, and Trustworthiness', *Ethics and Information Technology* 26: Art. 41.
32. Ho, Y.; Chen, B.-Y.; Li, C.-M. 2023. 'Thinking More Wisely: Using the Socratic Method to Develop Critical Thinking Skills Amongst Healthcare Students', *BMC Medical Education* 23: Art. 173.
33. Hugenroth, P.; Danry, V.; Maes, P. 2026. 'Critical Inker: Scaffolding Critical Thinking in AI-assisted Writing Through Socratic Questioning', in *Proceedings of CHI 2026*. New York: ACM.
34. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Bang, Y. J.; Madotto, A.; Fung, P. 2023. 'Survey of Hallucination in Natural Language Generation', *ACM Computing Surveys* 55(12): 1–38.
35. Krzanowski, R.; Pasterczyk, P. 2026. 'The Epistemological AI Turn: From JTB to Knowledges', *Philosophies* 11(1): 1–18.
36. Li, C. 2025. 'AI as a Socratic DIALOGUE PARTNER: An Intervention Study on Enhancing Students' Critical Thinking Skills', *AI in Education and Society* 1(1): 17–34.
37. Löhrer, G. 2022. 'Intellectual Modesty in Socratic Wisdom: Problems of Epistemic Logic and an Intuitionist Solution', *Logos & Episteme* 13(4): 411–429.
38. López-Goyez, J. P.; Romero, C.; Ventura, S. 2026. 'An Exploratory Study of a Generative AI-based Intelligent Tutoring System Using a Multi-agent Architecture in Higher Education', *Computers and Education: Artificial Intelligence* 8: 100345.
39. Makrakis, V. 2026. 'Aristotle and AI in Education: Virtue, Wisdom, Human Flourishing and the Common Good', in *Encyclopedia of Education and Information Technologies*. Cham: Springer.
40. Mitchell, M.; Krakauer, D. C. 2023. 'The Debate over Understanding in AI's Large Language Models', *Proceedings of the National Academy of Sciences* 120(13). Available at: https://doi.org/10.1073/pnas.2215907120
41. Mugleston, J. D.; Truong, V. H.; Kuang, C.; Sibiya, L.; Myung, J. 2025. 'Epistemology in the Age of Large Language Models', *Philosophies* 10(1): Art. 3.
42. Naser, M.; Naser, Z. 2025. 'Epistemic Virtues and Ethics of Explanation in Machine Learning', *AI and Ethics*. Available at: https://doi.org/10.1007/s43681-025-00869-4
43. Paul, R.; Elder, L. 2006. *The Thinker's Guide to the Art of Socratic Questioning*. Tomales: Foundation for Critical Thinking.
44. Qi, J.; Xu, Z.; Shen, Y.; Liu, M.; Jin, D.; Wang, Q.; Huang, L. 2023. 'The Art of Socratic Questioning: Recursive Thinking with Large Language Models', in *Proceedings of EMNLP 2023*. Stroudsburg: ACL, 4177–4199.

45. Sullins, J. P. 2019. 'The Role of Consciousness and Artificial Phronēsis in AI Ethical Reasoning', in *Proceedings of the AAAI Spring Symposium on AI Ethics*. Stanford: AAAI Press.
46. Telenti, A.; Auli, M.; Hie, B. L.; Maher, C. 2024. 'Large Language Models for Science and Medicine', *European Journal of Clinical Investigation* 54(6): e14183.
47. Tsai, C.-H.; Ku, H.-L. 2024. 'Why AI may Undermine Phronesis and What to do About it', *AI and Ethics*. Available at: https://doi.org/10.1007/s43681-024-00617-0
48. Tuozzo, T. M. 2011. *Plato's Charmides: Positive Elenchus in a 'Socratic' Dialogue*. Cambridge: Cambridge University Press.
49. Vallor, S. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford: Oxford University Press.
50. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhinet, I. 2017. 'Attention is All You Need', in *Advances in Neural Information Processing Systems 30*. Red Hook: Curran, 5998–6008.

AISTĖ DIRŽYTĖ, ALEKSANDRAS PATAPAS

# Ar dirbtinis intelektas gali būti sokratiškas?

*Santrauka*

Straipsnyje supažindinama su sokratiškojo intelekto (SI) koncepcija ir analizuojama, kaip didieji kalbos modeliai (DKM) yra susiję su dirbtinio ir žmogiškojo mokymosi procesais. Siekiant suvaldyti klaidų ir šališkumo grėsmes, siūlomas „pateisinamai teisingo algoritmo" principas, paremtas sokratiškuoju klausinėjimu, kuris reikalauja, kad atsakymai būtų pagrįsti, kuo tikslesni ir skaidrūs.

SI koncepcija siūlo holistinę dirbtinių ir žmogiškųjų mokymosi procesų perspektyvą, pabrėždama, kad dirbtiniai ir žmogiškieji agentai mokosi stebėdami, sąveikaudami ir reflektuodami sąveikas bei stebėjimo duomenis. Remiantis kitų autorių darbais, teigiama, kad mokymosi rezultatus lemia metakognityviniai gebėjimai ir pastiprinimo mechanizmai, o sokratiškasis intelektas gali būti apibrėžtas kaip apimantis pažintinį smalsumą, kritinį savirefleksyvumą, prielaidų analizę, gebėjimą palaikyti dialogą (taip pat ir vidinį) bei intelektinį nuolankumą.

Straipsnyje keliama prielaida, kad integruotas į išmaniąsias mokymo sistemas ar didžiuosius kalbos modelius, sokratiškasis intelektas gali padėti mokytis, skatindamas kritiškai mąstyti: analizuoti ir tikrinti prielaidas, informaciją, išvadas. SI gali būti pritaikytas paties dirbtinio intelekto algoritmų procesams, kad sistemos dinamiškai tikrintų savo sprendimus ir nuosekliai tobulėtų. Straipsnyje teigiama, kad taip galima sumažinti „haliucinacijų" bei šališkų išvadų rizikas, nes patys modeliai įgalinami savireguliuotis. Ateities tyrimai turėtų empiriškai patikrinti SI efektyvumą skirtinguose kontekstuose ir įvertinti, kokiomis sąlygomis SI iš tiesų gali pagerinti mokymosi rezultatus bei dirbtinio intelekto patikimumą.

**Reikšminiai žodžiai:** sokratiškasis metodas, didieji kalbiniai modeliai (DKM), intelektinis nuolankumas, *phronesis*, dirbtinio intelekto epistemologija