Can awareness be fairly evaluated across artificial systems?

Artificial systems are increasingly important in our daily lives, and their influence is only set to expand further. Current applications range from credit scoring and medical diagnosis to potential uses in wildfire response and goods delivery. In response to this growing ubiquity of artificial intelligence, philosophical debates have increasingly centred on questions of AI consciousness and moral status. Some authors explore the future possibility of AI consciousness by extending prominent theories of (human) consciousness (Butlin et al., 2023; Chalmers, 2023), while others argue that conscious AI is inevitable (Blum & Blum, 2025). By contrast, critics question whether specific architectural, developmental, or evolutionary conditions (Aru et al., 2023)—or even a biological substrate—are necessary preconditions (Block, 2025; Seth, 2026). Yet despite the intensity of this debate, there has been remarkably little progress on how such properties could, in practice, be evaluated.

Unfortunately, time may not be on our side when it comes to addressing these open questions (Schwitzgebel, 2025; Shevlin, 2024). For one, there are good reasons to be concerned that current discussions about the possibility of machine consciousness already influence public opinion (Deroy, 2023) and shape interactions with artificial systems (Colombatto et al., 2025). Moreover, as AI systems become increasingly integrated into everyday life and deployed in high-stakes contexts where they need to coordinate or collaborate with other agents (whether human or artificial), we require strategies to evaluate and engineer their capacities to ensure responsible development and deployment (Floridi, 2018). This raises a fundamental question: Is consciousness the right conceptual tool for this task, or is a better alternative available?

In recent work, awareness has increasingly been explored as a more neutral and tractable alternative to consciousness for guiding the development, oversight, and evaluation of artificial systems in AI and robotics research (Bacciu et al., 2025; Della Santina et al., 2025; Deroy et al., 2024; Evers et al., 2025). In this context, awareness refers to a system's ability to process, store, and utilise information in the service of goal-directed action (Lee et al., 2026), characterising context-sensitive information processing—specifically, the capacities artificial systems have to selectively register and respond to environmental, social, or internal changes. Crucially, this notion of awareness avoids premature or problematic attributions of consciousness to machines while redirecting design goals away from engineering conscious AI toward a more functional understanding of system capacities. As such, awareness retains

explanatory force while offering a practically valuable framework for AI design and evaluation.

If awareness is to play a functional role as a property of artificial systems, it must be susceptible to structured evaluation. This paper introduces a practical method for evaluating awareness across a diverse range of artificial systems. The proposed framework is guided by four desiderata for the fair and structured evaluation and comparison of such systems. First, evaluation must be domain sensitive. Although Artificial General Intelligence remains a driving goal (Goertzel, 2014), most current systems are specialised—a potentially preferable approach given concerns about sustainability and control (Deroy et al., 2024). Evaluation must therefore account for each system's operational domain. Second, it must be multidimensional, capturing both differences and overlaps in system capacities (cf. Birch et al., 2016 and Evers et al., 2025). Third, it must be deployable at different scales, allowing assessment at different levels of organisation, particularly in modular, distributed, or multi-agent systems such as swarm robotics (Brambilla et al., 2013). Fourth, it must predict task performance while generalising at the level of abilities, tracking underlying competences rather than isolated performances to enable principled and fair inter-system comparisons (Firestone, 2020).

Given these four desiderata, a structured approach is outlined for evaluating and comparing awareness profiles across artificial systems with differing architectures, scales, and operational domains. This approach comprises three interconnected elements: (i) dimensions of awareness that categorize distinct informational domains, (ii) action-perception abilities through which systems demonstrate awareness of these domains, and (iii) evaluative tasks designed to assess such abilities systematically.

Five key dimensions of awareness are identified that cut across these elements. Spatial awareness concerns a system’s abilities to detect, differentiate, and exploit spatial relations, such as distance, direction, or proximity. Temporal awareness involves detecting, differentiating and exploiting temporal relations, such as duration, continuity, and succession. Self-awareness pertains to monitoring information about the system's own (physical) states. Metacognitive awareness encompasses monitoring and evaluating one's own processing, uncertainty, and performance. Finally, agentive awareness relates to information about goals, intentions, and the causal relationship between actions and outcomes. Each dimension represents a distinct informational domain that can be leveraged in action-perception

couplings, allowing for granular assessment of what a system is aware of and how this awareness manifests in behaviour.

This framework addresses a central challenge in contemporary debates about AI consciousness: how to develop principled methods for evaluating the capacities of artificial systems. By repositioning the focus from consciousness to awareness, this approach offers a pragmatic lens that shifts the conversation away from whether artificial intelligence can replicate all human capacities. Instead, it directs attention toward identifying what artificial systems actually need to accomplish within specific domains, and which awareness capacities are required for systems to succeed at those tasks. This shift enables more tractable evaluation while remaining sensitive to the operational realities of artificial systems.

References

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences*, *46*(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

Bacciu, D., Ambriola, V., Bahrami, B., Ceni, A., Cossu, A., De Caro, V., Della Santina, C., Deroy, O., Gallicchio, C., Guidotti, R., Hauert, S., Jones, S., Karpus, J., Lee, S., Liu, J., Lomonaco, V., Meertens, N., Milner, E., Monreale, A., … Stölzle, M. (2024). EMERGE - Emergent Awareness from Minimal Collectives. In C. Secchi & L. Marconi (Eds), *European Robotics Forum 2024* (Vol. 32, pp. 87–91). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76424-0_16

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007

Block, N. (2025). Can only meat machines be conscious? *Trends in Cognitive Sciences*, S1364661325002347. https://doi.org/10.1016/j.tics.2025.08.009

Blum, L., & Blum, M. (2024). *AI Consciousness is Inevitable: A Theoretical Computer Science Perspective* (Version 14). arXiv. https://doi.org/10.48550/ARXIV.2403.17101

Brambilla, M., Ferrante, E., Birattari, M., & Dorigo, M. (2013). Swarm robotics: A review from the swarm engineering perspective. *Swarm Intelligence*, *7*(1), 1–41. https://doi.org/10.1007/s11721-012-0075-2

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2308.08708

Chalmers, D. J. (2023). *Could a Large Language Model be Conscious?* https://doi.org/10.48550/ARXIV.2303.07103

Colombatto, C., Birch, J., & Fleming, S. M. (2025). The influence of mental state attributions on trust in large language models. *Communications Psychology*, *3*(1), 84. https://doi.org/10.1038/s44271-025-00262-1

Della Santina, C., Corbato, C. H., Sisman, B., Leiva, L. A., Arapakis, I., Vakalellis, M., Vanderdonckt, J., D'Haro, L. F., Manzi, G., Becchio, C., Elamrani, A., Alirezaei, M., Castellano, G., Dimarogonas, D. V., Ghosh, A., Haesaert, S., Soudjani, S., Stroeve, S.,

Verschure, P., … Sierra, C. (2024). Awareness in Robotics: An Early Perspective from the Viewpoint of the EIC Pathfinder Challenge “Awareness Inside”. In C. Secchi & L. Marconi (Eds), *European Robotics Forum 2024* (Vol. 32, pp. 108–113). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76424-0_20

Deroy, O. (2023). The Ethics of Terminology: Can We Use Human Terms to Describe AI? *Topoi*, *42*(3), 881–889. https://doi.org/10.1007/s11245-023-09934-1

Deroy, O., Bacciu, D., Bahrami, B., Della Santina, C., & Hauert, S. (2024). Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General Intelligence and Artificial Consciousness. *Advanced Intelligent Systems*, *6*(10), 2300740. https://doi.org/10.1002/aisy.202300740

Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P. F. M. J., & Khamassi, M. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. *Physics of Life Reviews*, *52*, 180–193. https://doi.org/10.1016/j.plrev.2025.01.002

Firestone, C. (2020). Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, *117*(43), 26562–26571. https://doi.org/10.1073/pnas.1905334117

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, *5*(1), 1–48. https://doi.org/10.2478/jagi-2014-0001

Lee, S., Meertens, N., Milner, E., & Hauert, S. (2026). A Framework for the Examination of Awareness in Artificial Systems. In A. Jiménez Rodríguez, R. Mestre, C. Chen, A. Mura, E. Barker, P. Verschure, & T. Prescott (Eds), *Biomimetic and Biohybrid Systems* (Vol. 15582, pp. 320–333). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-07448-5_27

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence. *Frontiers in Ecology and Evolution*, *9*, 806283. https://doi.org/10.3389/fevo.2021.806283

Schwitzgebel, E. (2025). *AI and Consciousness* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2510.09858

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42. https://doi.org/10.1017/S0140525X25000032

Shevlin, H. (2024). Consciousness, Machines, and Moral Status. In A. Strasser (Ed.), *Humans and smart machines as partners in thought*. Xenomoi.