Conscious AI and Human Communication

Thomas Herrmann
Institut für Arbeitswissenschaft, University of Bochum, Germany, thomas.herrmann@rub.de

Consciousness depends on communication between individuals who consider each other as alter ego. It seems to be questionable whether AI agents should be considered as a set of separated conscious individuals that cannot not directly be connected to each other. However, individualized consciousness can serve as a useful metaphor for designing human-computer interaction.

CCS
Human-centered computing, Human computer interaction (HCI), Interaction paradigms

KEYWORDS
Human-centered AI, Conscious AI, communication, human-computer interaction, socio-technical design

1 Background of Consciousness

Based on the following approaches:

- Symbolic Interactionism [1]
- Reciprocity of Perspectives [2]
- Epistemology according to Maturana and Varela [3]
- Theory of social systems [4]

we assume that consciousness is a phenomenon that initially manifests itself in interpersonal communication. People experience communication that highlights the difference between the inner and outer worlds as well as between inner and outer actions [5]. Inner actions, like thinking through a certain topic, is only perceptible by the actor himself. The consciousness of oneself and the continuous "Stream of Thoughts" [6] can only be experienced by another person through communication. We, as humans, assume that another person represents an “alter ego” [7] that is similar to us in many ways and accordingly performs inner actions and possesses consciousness but is a different individual. This inner world and the associated consciousness are fundamentally inaccessible from the outside; communication about inner states is fundamentally fallible [5] meaning we can never be sure that our communications, including descriptions about our consciousness, are understood correctly. We perceive other people as individuals who each have their own consciousness accessible only to themselves, and we can learn of this consciousness only by these individuals talking about themselves. This is a fundamental condition of the social interaction between human individuals, which is established in the interplay between fundamental independence (thought is free) and mutual social interdependence.

In many representations, whether in novels or films, but also in scientific concepts or application scenarios, AI agents are portrayed as phenomena comparable to distinct human individuals. However, this assumption is obsolete in view of the internet and mobile communication capabilities. In principle, AI software has the possibility to communicate with other AI software – limitations can be set through technical design but are also, in principle, reversible. A community of AI individuals in the form of fundamentally separate AI agents having similar experiences to human communities is unlikely. The possibilities for inspecting or adapting internal processes of AI agents via interfaces seem indispensable. Continuous data exchange among AI agents or with central servers appears essential. As users of technology, we are so accustomed to the possibility of constant data exchange between devices such as smartphones that it is hard to imagine these possibilities should not be available for AI agents. It will therefore likely be that such agents could potentially interact directly, not only symbolically, with each other in the sense that they share copies of their internal states and processes as data, rather than by communicational descriptions of inner actions as

Mensch und Computer 2024 – Workshopband, Gesellschaft für Informatik e.V., September 1–4, 2024, Karlsruhe, Germany

https://doi.org/10.18420/muc2024-mci-ws12-382

is the case with humans. The socialization experiences of AI agents among themselves—so here is the hypothesis—will therefore present themselves fundamentally differently from the socialization experience of humans. Only for the purpose of scientific experiments, one might want to have AI agents act in artificial isolation.

Consciousness in AI, therefore, is not something attributed to individual AI agents but is a global phenomenon of globally networked AI. For the purpose of continuous development and quality assurance, it is beneficial if AI agents learn from each other with respect to task and error handling. Thus, if the metaphor of consciousness should make sense in the case of AI, it will not be reasonable to attribute consciousness to single AI agents but to the network of interacting AI instances and interfaces as a whole within socio-technical processes.

2 Human-Computer Interaction and conscious AI

Regarding interaction scenarios, we will be able to distinguish two levels:

a) The AI interacts with others as if it has an individual consciousness and therefore has inaccessible internal states. This can be helpful for scientific experiments or for certain socio-technical processes of psychological care or social support for humans. In such a scenario, it may also be conceivable that one might have to persuade an AI agent to do something specific—e.g., for training purposes where AI is used as a sparring partner [8]. However, in general, the application of technology should be designed in a way that delegating tasks to the AI involves as little effort as possible.

b) There is the possibility of directly delegating tasks to AI in the sense of exercising external control. For example, a coach in an executive training program might configure an AI agent to act as a difficult-to-motivate employee towards a trainee. The coach should not have to persuade the AI agent to play this role. For the purpose of quality assurance, AI software should be able to be directly inspected, for example through exploration and intervention. With intervention [9] temporary changes are possible to directly control AI or to temporarily change parameters. Small changes can also serve exploration to understand or test processes in the AI.

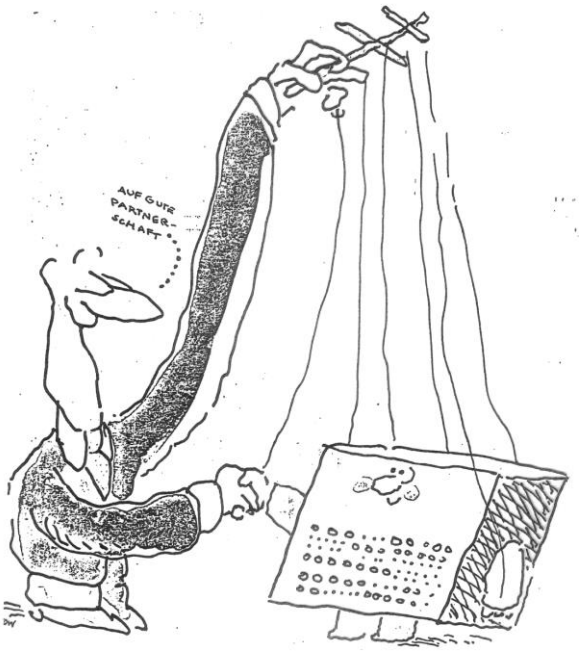

Figure 1: The tension between partnership and tool-perspective [10] (speech bubble says “on good partnership)

Figure 1 represents these two levels in a visual illustration. The transition between these two levels will ultimately be determined by socio-technical design, which includes technical functionality, the development of suitable organizational practices and the appropriate assigning of roles to humans and AI agents [11].

References

[1] G. H. Mead, *Mind, Self and Society*. London: University of Chicago Press 3 Aufl 1967, 1934.
[2] A. Schutz, "Common-sense and scientific interpretation of human action," in *Collected papers I: The problem of social reality*, Springer, 1962, pp. 3–47.
[3] H. Maturana and F. Varela, *Der Baum der Erkenntnis*. Bern, München, Wien: Scherz, 1987.
[4] N. Luhmann, *Social Systems*. California: Stanford University Press, 1995.
[5] G. Ungeheuer and others, "Vor-Urteile über Sprechen, Mitteilen, Verstehen," *Kommunikationstheoretische Schriften*, vol. 1, pp. 229–338, 1982.
[6] W. James, "The stream of consciousness," *Psychology*, pp. 151–175, 1892.
[7] A. Schütz, *Der sinnhafte Aufbau der sozialen Welt*. Suhrkamp, 1974.
[8] T. Herrmann, "Evolution of interaction-free usage in the wake of AI," *i-com*, vol. 0, no. 0, Jun. 2024, doi: 10.1515/icom-2024-0005.
[9] A. Schmidt and T. Herrmann, "Intervention user interfaces: a new interaction paradigm for automated systems," *interactions*, vol. 24, no. 5, pp. 40–45, 2017.
[10] T. Herrmann, *Rationalität und Irrationalität in der Mensch-Computer Interaktion (unpublished Master Thesis)*. Bonn: University of Bonn, 1983. [Online]. Available: DOI: 10.13140/RG.2.2.35273.21607
[11] I. Jahnke, C. Ritterskamp, and T. Herrmann, "Sociotechnical Roles for Sociotechnical Systems: a perspective from social and computer science," in *AAAi Fall Symposium Proceedings*, 2005, pp. 68–75.