Paper prepared for
THE SHERRY TURKLE MIRACLE
online workshop, May 27, 2021, organized by the
Committee for Communication and Media Theory,
Hungarian Academy of Sciences.

Judit Szalai

Understanding and Predicting the Behaviour of Artificial Agents

One of the leading themes of Sherry Turkle's more recent work (*Reclaiming Conversation: The Power of Talk in a Digital Age*; *Alone Together: Why We Expect More from Technology and Less from Each Other*) is the drive to have sustained communicative and emotional connection with fellow-humans, the fulfilment of which is partially undermined by the use of current digital technologies. She identifies novel tendencies in relating to ourselves and to others, altered social routines that home computers, social media platforms, android phones and other smart devices enable and encourage. Current technologies serve as channels or outlets for discharging our emotional-interactive needs in ways that often modify our perception of these needs themselves or compel us to imaginarily bend our communicative possibilities.

One such strategy is extensively examined in *The Second Self*, where Turkle discusses our tendency to invest machines with psychological attributes. The behaviour of even well-informed, sophisticated users towards ELIZA, an early computer psychotherapist, is observed to be markedly anthropomorphizing. Being aware of its lack of emotional capacities and very limited cognitive repertoire, users "went out of their way to ask questions in a form that they believed would provoke a lifelike response", in an attempt to "maintain the illusion that ELIZA was able to respond to them". The inclination to understand artificial systems' behaviour in terms of human psychology, which Turkle also amply demonstrates in the case of children, has

been confirmed since by anthropological studies (e.g., B. Chun and H. Knight: “The Robot Makers…”).

This tendency will no doubt be enhanced by the increasingly broadening functions of humanoid and non-humanoid robots, especially those with agentic (self-learning and decision-making) properties. When it comes to other agents, human or non-human, we are bound to try to understand, explain and predict their behaviour. People are folk psychological reasoners. Folk psychology is not just a matter of convenience, it is ineliminable. While we are likely to find it easier to connect to AI-based systems with humanoid features, being unable to make sense of the actions of an autonomous car or a “sex robot” could equally have unwelcome consequences. Systematically lacking such folk psychologies while having to count with non-human agents making decisions and acting around, and for, us could be paralyzing or lead to rather suboptimal behaviour.

Our anthropomorpizing tendencies can thoroughly lead us astray, however. It would be a grave mistake to use the same folk psychology for artificial agents we do with our conspecifics. Artificial agents not just lack certain human functions, especially affective ones. Their purely “cognitive” functions are also qualitatively different, e.g., the “reasoning” used in machine learning is also largely inaccessible to human thinking. With the transition of artificial systems from being instrument-like to agentic, the chances and stakes of understanding and prediction equally change. As Kaplan describes it, in the past, programmers fully understood the steps required for a computer to accomplish a task and then wrote “a program that, in effect, cause[d] the machine to simulate these steps precisely” (J. Kaplan, *Humans Need Not Apply*). “Synthetic intellects”, in contrast, “are not programmed in the conventional sense… where they wind up is unpredictable and not under their creator’s control.”

Relying on prediction and interaction is not made possible, as in human cases, by the fact that, being members of the same species, we share the same kind of physical-biological body and mental make-up. The radical qualitative difference of AI-based agents’ mental-like processes is only one issue, though. Another factor that will complicate what is often called “human-robot-interaction” is the hetero-

genity of the set of AI-based systems. The therapeutic baby seal robot PARO, with a limited range of actions, is very different from neural networks with biological cells extracted from mouse embryos.

Since analogous thinking relying on a theory of the human mind is not an option with non-human agents, we need to find appropriate epistemic channels to secure some measure of predictability and explainability in interaction with artificial entities. Just as the need for the regulation of the capacities and functions of AI-based systems has recently been clearly realized (see, e.g., the work of the European Union's Higher-Level Expert Group on AI), having, for instance, the external design of the AI system reflect those capacities and functions to help understanding and prediction, and thereby inform interaction, also seems inevitable.