Anticipatory behavior for safety in human-autonomous agent interactions

Automation is the application of technologies, processes, and robotics to tasks which ordinarily would have been performed by humans (Parasuraman et al., 2000). Technologies in autonomous agents can communicate intent with humans and other autonomous agents, they can read the situation around them and initiate actions and change the operating rules through adaptive behavior when necessary. Likewise, human beings have used mental models and transactive memory systems for millennia to anticipate and predict intentions thus ensuring adaptive behaviors to their environment (Cannon-Bowers et al., 1993; Endsley, 1995). We are currently witnessing autonomous agents possessing the ability to interact with human agents in domains usually reserved for interaction among human agents. The rules of such interaction are not exclusively based on human agents' possessing control over the autonomous agents but rather both are equal status participants in these arenas. Indeed, there seems to be the suggestion, albeit not yet realized, that autonomous agents will soon replace human agents in these spaces. This seems far-fetched and a realistic scenario will involve human agents and autonomous agents acting together in these shared spaces. Among the most important consideration in this interactive space is how human agents and autonomous agents learn to anticipate each other's intentions and behavior to ensure adaptive co-action. Thus, in the vein proposed by the human centered artificial intelligence (HCAI) framework (Shneiderman, 2020), autonomous agents will not replace human agents. Rather, to assure safe, trustworthy, and reliable systems, high level of control must be ceded to both human agents and autonomous agents. This means that in some tasks, human agents possess high levels of control and on some other tasks, autonomous agents possess high levels of control. Thus, one must consider various objectives for which it is functional to have humans have high or low degree of control, and vice versa for autonomous systems. This framework thus argues for design that keeps humans in the action performance loop rather than outside the action performance loop when automation is high. The consideration of human agent and autonomous agent coaction and control is very important since presently, autonomous agents may struggle with anticipating intentions and adapting to unexpected situations. These are features of the environment which are especially important in dynamic environments where safe coaction depends on the ability to anticipate intention and engage in adaptive behavior. While human agents have not necessarily being perfect at resolving those demands, we have a good understanding of how humans anticipate each other's intentions and behaviors based on processes underscored by bio-cognitive processes and social conventions. Anticipation is fundamentally an awareness of what events are likely to follow in the future given the current condition (Castiello, 2003). It requires an amalgamation of a subject's internal state and inference of an observer's external signals. Humans use social cues such as eye contact and bodily positioning to signal intent and to infer intention. In the social arena, humans have an internal mental representation of each other's behavior and can thus simulate their co-actors ongoing behavior thus anticipation depends more on attention to the ongoing behavior (Bisio et al., 2014; Castiello, 2003). The question that arises is what are the equivalent processes in autonomous agents? How can these processes impact co-action in collaborative spaces (work setting, social settings, transport) where humans and autonomous systems have to interact. The purpose of this proposed special session during the ESREL2025 conference is to invite guest speakers and researchers to explore the theoretical, empirical, and practical dimensions of this issue to contribute to the understanding and adoption of autonomous systems in our evolving societies.

Provisional List of Guest Speakers

Dr. Prosper Kwei-Narh (Session Convener)

Snr Research Scientist, Institute for Energy Technology

Prof. Jan Ketil Arnulf

BI Norwegian Business School

++ more guests to be confirmed

Reference

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *Plos One*, *9*(8), e106172.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In N. J. Castellan (Ed.), *Individual and group decision making* (pp. 221-246). Lawrence Erlbaum.

Castiello, U. (2003). Understanding other people's actions: intention and attention. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 416.

Endsley, M. R. (1995). A Taxonomy of Situation Awareness Errors. In R. Fuller, N. Johnston, & N. McDonald (Eds.), *Human factors in aviation operations* (Vol. 3, pp. 287-292). Ashgate.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). *A model for types and levels of human interaction with automation* [286-297]. New York, NY :.

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118