Joint agency in human-machine interaction: How to design more cooperative agents?

Labex topics: Modeling of humans & Interfaces and Interaction with humans

Director: Mohamed Chetouani (ISIR); Co-directors: Bruno Berberian (ONERA), Ouriel Grynszpan (ISIR)

Doctoral School: SMAER

The use of automation technology in process industries tends to steadily increase. This happens since automation technology offers efficiency and stable control at the same time as it makes the control room operators' job easier in many ways. When new automation is introduced into a system or when there is an increase in the autonomy of automated systems, developers often assume that adding "automation" is a simple substitution of a machine activity for human activity. However, empirical data on the relationship of people and technology suggest that is not the case and that traditional automation has many negative performance and safety consequences including human vigilance decrements (e.g., lack of operator sensitivity to signals) (Billings, 1991), complacency (e.g., over trust in highly reliable computer control) (Parasuraman, Molloy, Singh, 1993) and loss of operator situation awareness (Endsley & Kiris, 1995). Although difficulties with automated systems have been identified for a long time, clear solutions are still missing to overcome them (Norman, 2010).

Adding or expanding the machine's role changes the cooperative architecture, changing the human's role, often in profound ways (Sarter, Woods, and Billings, 1997). Creating partially autonomous machine agents is, in part, like adding a new team member and creates new coordination demands for the human operator. Where designers really need guidance today is **how to support the coordination between people and automation**, not only in foreseeable standard situations, but also during novel, unexpected circumstances. As pointed by Norman (1990), the main problem with automation is not the presence of automation per se, but rather its inappropriate design. In this sense, making automation a collaborative agent (Klein & al, 2004) becomes a first concern. The first goal of this PhD is to help automation designers to achieve this goal.

The design of collaborative agent has raised a particular interest during the last years (Christoffersen & Woods, 2000; Klein & al, 2004; Hoc, 2007; Dragan, Lee, & Srinivasa, 2013). In this PhD, we seek to contribute to this effort by introducing recent insight about how humans understand and control joint action. Indeed, we can assume that operators interpret the intentions and the outcomes' actions of a system with their own "cognitive toolkit". Thus, understanding how this "cognitive toolkit" works could be relevant to propose design principles for potentially controllable/collaborative systems.

The mechanisms underlying the experience of intentional causation and the sense of control of our own actions are the first concern of the science of agency. Gallagher (2000) defined agency as "the sense that I am the one who is causing or generating an action". Put differently, agency corresponds to our capacity to make things happens, to change the world thorough our action. Although, the mental processes contributing to the sense of agency are not fully understood at this time, a classical approach proposes that we derive a sense of being the agent for our own actions by a cognitive mechanism that computes the discrepancies between the predicted consequences of our own actions and the actual consequences of these actions, similarly to action control models (see Blakemore, Wolpert, & Frith, 2002). Interestingly, Pacherie (2012) argued that the different mechanisms underlying the sense of agency for individual actions are of the same kind than those underlying the sense of agency that one experiences when engaged in joint action. That is, the sense of agency in joint action is based on the same principle of congruence between predicted and actual outcomes.

We can imagine that in the same manner than when two people work together, the supervisors must be able to predict automated systems' actions and their outcomes in order to facilitate the cooperation between them and built a "we-agency" (or joint agency). This proposition echoes that of Norman (1990) when he assumed that what is needed is continual feedback about the state of the system, in a normal natural way, much in a similar manner to human participants in a joint problem-solving activity who discuss the issues among themselves. However, empirical data (Wohlschläger, Haggard, Gesierich, & Prinz, 2003; Obhi & Hall,

2011) seems to indicate that joint agency is more difficult to develop when we interact with computers. Explaining these difficulties and suggesting design recommendation to overcome them is a major and exciting issue.

In this sense, the goal of this project is to explore how to design automation technology to make it a joint agent. In this sense, we will try to understand (1) how our sense of agency for actions differs when collaborating with computer vs. human partners, (2) how to design artificial agents which allow the "we-agency" to develop (increase the humanness of the artificial agent by using an avatar, feedback regarding artificial agent's intention), (3) what are the change involved by the "we-agency" from the human operator point of view (performance, acceptability, confidence, …).

PhD program

- Literature review:
  a. Joint action and cooperative agent;
  b. Model of agency and Joint agency mechanisms (Factors contributing to the sense of agency, …)
- Series of experimentation to compare sense of agency in case of human-human interaction and human-robot interaction. Factors manipulation (humanness of the artificial agent with the use of avatars, feedback regarding artificial agent's intention) to optimize the sense of agency in case of human-robot interaction.
- Elaboration of an artificial agent model in view to develop this mutual sense of agency (we-agency).
- Experimentation for Model evaluation

References

Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines*.

Blakemore, S. J., Wolpert, D. M., Frith, C. D. (2002): Abnormalities in the awareness of action. *Trends in cognitive sciences, 6(6),* 237-242

Christoffersen, K., & Woods, D. D. (2002). How to make automated systems team players. Advances in human performance and cognitive engineering research, 2, 1-12.

Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference, pp. 301-308. IEEE.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37(2),* 381-394.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences, 4(1),* 14-21.

Hoc, J. M. (2007). *Human and automation: a matter of cooperation*. In HUMAN 07 (pp. 277-285).

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems, 19(6),* 91-95.

Norman, D. A. (1990). The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation'. *Philosophical Transactions of the Royal Society B: Biological Sciences, 327(1241),* 585-593.

Norman, D. A. (2010). *Living with complexity.* Mit Press.

Obhi, S. S., & Hall, P. (2011). Sense of agency and intentional binding in joint action. *Experimental brain research, 211*(3-4), 655-662.

Pacherie, E. (2012). *The phenomenology of joint action: Self-agency vs. joint-agency*. Joint attention: New developments, 343-389.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology, 3*(1), 1-23.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of human factors and ergonomics, 2,* 1926-1943.

Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self-and other-generated actions. *Psychological Science, 14*(6), 586-591.