Le projet InStance : l'intentionnalité dans les interactions humain-robot sous le prisme des neurosciences cognitives et computationnelles

Marwen Belkaid et Agnieszka Wykowska
Istituto Italiano di Tecnologia (IIT), Gênes, Italie.

Quelle attitude les humains adoptent-ils en interagissant avec les objets artificiels qu'ils créent ? Dans sa théorie, Dennett (1987) a proposé trois modes utilisés selon la nature et la complexité des objets observés. L'attitude physique (*physical stance*) permet d'expliquer le comportement d'un objet en fonction de ses propriétés mécaniques et de notre connaissance des lois de la physique. L'attitude conceptive (*design stance*) se base quant à elle sur la connaissance que l'on a du fonctionnement pour lequel l'objet a été conçu. Enfin, l'*attitude intentionnelle* (*intentional stance*) traite l'entité observée comme un agent autonome et en interprète le comportement sur la base d'états mentaux attribués (e.g. croyances, désirs, intentions, émotions).

Une lecture simpliste de ce cadre théorique pourrait prédire que nous adoptions systématiquement l'attitude conceptive face aux objets conçus par l'humain. Ce serait sans compter sur notre tendance naturelle à l'anthropomorphisation, et en particulier à prêter des états mentaux, même ponctuellement, à des objets qui n'ont en pas. L'un des facteurs impliqués dans ce processus est le degré de connaissance de l'objet observé : plus il est complexe, plus l'on est susceptible de lui prêter une forme d'agentivité (Waytz et al., 2010). Or, les robots ont la particularité d'être parmi les objets les plus complexes conçus par l'humain pouvant à terme entrer dans la vie quotidienne d'un grand nombre de personnes. Il est donc crucial de se poser des questions sur leur perception par des humains en majorité non-experts et ayant une faible connaissance de leur fonctionnement (Wiese et al., 2017).

Le projet ERC InStance vise justement à étudier les conditions dans lesquelles les humains pourraient adopter une attitude intentionnelle en interagissant avec les robots (voir https://www.instanceproject.eu/). C'est un projet interdisciplinaire avec approche centrée sur les neurosciences cognitives et computationnelles. Ainsi, nous mettons en places des protocoles expérimentaux permettant d'étudier le comportement et l'activité neuronale des participants dans des interactions humain-robot (IHR). Par exemple, Kompatsiari et collègues (2018) ont pu répliquer l'effet de *gaze cueing* (indication par le regard) ainsi que ses corrélats neuronaux (composantes P1/N1 du potentiel évoqué EEG/ERP) dans des interactions avec le robot humanoïde iCub.

Par ailleurs, Marchesi et collègues (2019) ont développé le questionnaire InStance permettant d'évaluer si les participants tendent à expliquer le comportement du robot avec des termes plutôt mécaniques ou mentaux dans une série de scénarios. Le résultats suggèrent que les humains peuvent dans certains contextes adopter l'attitude intentionnelle envers un robot et lui prêter des états mentaux. Ces deux études illustrent l'intérêt qu'il y à mobiliser un cadre théorique et méthodologique issu de la psychologie expérimentale et des neurosciences cognitives dans le but de quantifier des facteurs clés dans les IHR. Dans des travaux futurs, l'un des objectifs est d'aller plus loin que les mesures explicites et subjectives telles que le questionnaire InStance et de chercher des mesures objectives et implicites (e.g. signaux EEG, pupillométrie) de l'attitude intentionnelle.

Sur la base de résultats expérimentaux obtenus jusqu'ici, nous avons également entamé un travail de modélisation computationnelle pour analyser le comportement des humains faces aux robots. Par exemple, nous utilisons des modèles de prise de décision pour tester des hypothèses sur les biais individuels ou encore sur la prise en compte des signaux sociaux dans la réalisation d'une tâche.

Combinés aux données neuronales, ces modèles computationnels peuvent donner lieu à des prédictions sur les processus sous-jacents et les substrats neuronaux impliqués. L'objectif est aussi que le travail de modélisation puisse informer le volet expérimental et guider la conception de protocoles futurs.

Références :

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Kompatsiari, K., Pérez-Osorio, J., De Tommaso, D., Metta, G., & Wykowska, A. (2018, October). Neuroscientifically-grounded research for improved human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3403-3408). IEEE.

Marchesi, S., Ghiglino, D., Ciardo, F., Baykara, E., & Wykowska, A. (2019). Do we adopt the Intentional Stance towards humanoid robots?. *Frontiers in psychology*, *10*, 450.

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, *19*(1), 58-62.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology*, *8*, 1663.

Financement :

Ce projet a reçu un financement du Conseil Européen de la Recherche (ERC) dans le cadre du Programme de Recherche et d'Innovation de l'Union Européenne Horizon 2020 (subvention accordée à AW, intitulée "InStance: Intentional Stance for Social Attunement." ERC starting grant, Grant Agreement No. 715058).