Liens entre confiance et acceptabilité dans un dispositif IA

Alexandre Agossah, Lucie Lévêque, Matthieu Perreira da Silva, Patrick Le Callet, Frédérique Krupa, Guillaume Deconde

Liens entre confiance et acceptabilité dans un dispositif IA

Relationship between trust and acceptability in AI

ALEXANDRE AGOSSAH, Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 60004; Digital Design Lab, L'École de Design Nantes Atlantique; Groupe Sigma, France

LUCIE LÉVÊQUE, MATTHIEU PERREIRA DA SILVA, and PATRICK LE CALLET, Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 60004, France

FRÉDÉRIQUE KRUPA, Digital Design Lab, L'École de Design Nantes Atlantique, France

GUILLAUME DECONDE, Groupe Sigma, France

Abstract: Several studies have presented trust as crucial to predict AI acceptability. We aim to involve statistical and social measures of trust, to confirm links between trust and acceptability, and between trust and emotional variation, with an experimental protocol inspired by [1]. Sixty participants are asked to estimate ages on portraits; then, an AI model suggests its own prediction based on facial features, along with its stated confidence. Participants can then keep or change their initial prediction. Three measures are used to quantify their confidence in the model: the proportion of agreements and changes, and fixation duration. Participants are also asked to complete a questionnaire to measure their trust and acceptance, and their facial expressions are recorded during the experiment to assess emotional variation. Therefore, this research allows to better understand links between trust, emotions, and acceptability in AI.

Key words: Artificial intelligence, Acceptability, Trust, Emotional variation.

Résumé : Diverses études présentent la confiance comme essentielle pour tirer profit des solutions d'IA, et comme prédictive de son acceptabilité. Notre étude a pour objectif de mobiliser des mesures statistiques et sociales de la confiance, de réaffirmer les liens entre confiance et acceptabilité, et entre confiance et variation émotionnelle. Cet abstract présente le protocole expérimental mis en place, inspiré par [1]. Des portraits sont présentés à 60 participants, devant estimer leur âge ; puis un modèle IA propose sa propre prédiction à partir de caractéristiques faciales, ainsi que sa confiance déclarée. Les participants ont alors le choix entre conserver leur prédiction initiale ou changer de réponse. Pour mesurer le comportement des participants, nous récoltons les proportions d'accords et de changements face aux prédictions du modèle, et nous enregistrons le temps des fixations oculaires et leurs expressions faciales. Les participants remplissent également un questionnaire pour mesurer leur confiance et acceptation dans le modèle. Cette recherche nous permet ainsi de mieux comprendre les liens entre confiance, variation émotionnelle, et acceptabilité en l'IA.

Mots-clés : Intelligence artificielle, Acceptabilité, Confiance, Variation émotionnelle.

RÉFÉRENCES

[1] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.