Learning about Other Persons' Character Traits Relies on Combining Reinforcement Learning with Representations of Trait Similarities

Koen Frolichs (k.frolichs@uke.de)
Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistraße 52
Hamburg 20246, Germany

Benjamin J. Kuper-Smith (b.kuper-smith@uke.de)
Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistraße 52
Hamburg 20246, Germany

Jan Gläscher (glaescher@uke.de)
Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistraße 52
Hamburg 20246, Germany

Gabriela Rosenblau (grosenblau@email.gwu.edu)
Autism and Neurodevelopmental Disorders Institute, George Washington University and Children's National Health System, 2115 G Street NW
Washington, DC 20052, USA

Christoph W. Korn[1] (c.korn@uke.de)
Institute for Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Martinistraße 52
Hamburg 20246, Germany

Abstract:

Humans often describe other persons (and themselves) in terms of abstract character traits. When getting to know a new person, they need to update their estimates of the other person across many different character traits. It is unclear how this learning process unfolds and how the relationship between diverse character traits are represented in brain activity. Here, we first showed in three behavioral studies that humans combine reinforcement learning with their knowledge about the correlations between traits when learning about other persons' character. Second, in two functional imaging studies the fine-grained similarities between character traits were represented in medial prefrontal cortex, in a region that has consistently been linked to thinking about other persons. Our findings thus suggest a possible learning mechanism for rather complex generalization across character traits according to their similarities, which seem to be related to the medial prefrontal cortex.

Keywords: reinforcement learning; representational similarity analyses; medial prefrontal cortex

Introduction

Everyday social interactions vary substantially. Therefore, humans often need to abstract away from the interaction at hand and employ generalized concepts for similar persons. Humans tend to describe persons in terms of abstract character traits. In many social interactions, humans learn continuously about each other's personality traits (e.g., how polite, helpful, and reliable is the other person?). Formal models that capture such complex social learning processes are currently lacking.

Reinforcement learning models successfully describe human learning in several social contexts and therefore constitute good candidates. However, we surmised here that reinforcement learning per se is not sufficient to account for learning about personality traits since humans may additionally rely on their knowledge about the distributions of different traits and the similarities between them (e.g., most people are moderately to very polite and polite persons tend to be helpful).

The first aim of the to-be-presented work was to use behavioural modelling to assess whether the relations between character traits guide learning. The second

[1] corresponding author

aim was to provide evidence for a representation of the relations between these traits in functional magnetic resonance imaging (fMRI) signals within the medial prefrontal cortex, a region commonly involved in social cognition and in abstract representations.

Methods

Part 1: Social Learning about Character Traits

In three behavioral studies, participants (n=36; n=41; n=23) were asked to consecutively predict how four other persons had previously rated themselves on a series of 60 trait words (see Figure 1 for a schematic of the task). After each prediction, participants received veridical feedback (but they never met the other persons).

The three studies used different sets of trait words. Distributions and similarities of all traits were derived from self-ratings of independent laboratory and online samples (total n=835). Participants received veridical and manipulated feedback. We compared basic non-learning models against various models derived from the standard Rescorla-Wagner framework.

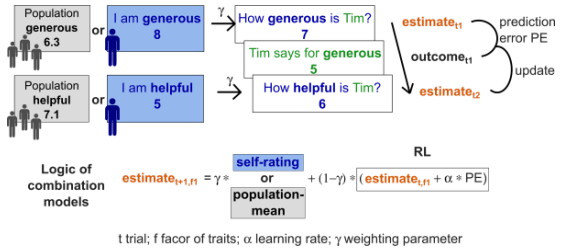

Figure 1: Social learning task and logic of the models used to describe participants' behavior.

Part 2: Representation of Character Traits

For the second aim, we applied representational similarity analyses in two fMRI studies, in which participants (n=27; n=30) were asked to think about themselves and to rate themselves on various character traits. Representational similarity analyses were conducted using both a priori regions of interest and searchlight procedures. Main analyses implemented correlations between traits. Control analyses used visual differences between trait words.

Results

Part 1: Social Learning about Character Traits

Bayesian model comparison showed that the winning behavioral models combined reinforcement learning with reliance on the distributions and similarities of the employed personality traits. Crucially, the winning models generalizes across traits according to their similarities. That is, when receiving information about a given trait the estimates of all other traits are updated according to how similar these are to the trait at hand. A comparison of the three studies showed that the implemented similarity metric does not depend on the range of traits to be learned about. Instead, it is crucial that feedback originates from veridical persons.

Part 2: Representation of Character Traits

Representational similarity analyses of fMRI data in non-learning tasks provided converging support for the plausibility of using the similarities between traits since these similarities were reflected within the medial prefrontal cortex, a region classically associated with reflecting about personality traits (Figure 2). As expected, control analyses showed that visual features of the trait words were related to visual cortex.

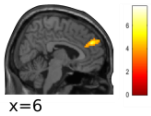

Figure 2: Results of representational similarity analyses linking the correlations between character traits to the medial prefrontal cortex.

Conclusions

Overall, our behavioral results indicate that variants of reinforcement learning algorithms which incorporate the similarities of character traits describe crucial aspects of the dynamics at play when persons interact each other. The imaging results suggest how these similarity structures could be stored in medial prefrontal cortex.

Acknowledgments

This work was funded by an Emmy Noether Research Group (392443797) awarded by the German Research Foundation (DFG).