COGNITIVE SCIENCE

Modelling theory of mind

The assumption of rationality is at the heart of action explanations. A Bayesian model of theory of mind, which explicitly relies on this assumption, can predict with high accuracy the inferences that people make about the mental states underlying others' actions.

Gergely Csibra

Why did the chicken cross the road? Whatever your preferred answer to this old joke, it will likely be an attempt to justify the chicken's action by outlining reasons that would make sense of this behaviour. As philosophers repeatedly pointed out^{1,2}, providing, or even finding, such reasons requires adopting the assumption that the to-be-explained behaviour is rational, in the sense of implementing a practical means to a certain end that is beneficial for the actor. People readily offer such explanations for their own actions as well as those of other humans³, but — as the joke demonstrates — also for the behaviours of animals, and even for the animated movement patterns of geometric figures⁴. In fact, even babies try to rationalize observed behaviours by assuming, and expecting, them to be cost-efficient (hence, rational) towards their goals⁵.

However, not all actions are successful, and many of them fail to meet the requirement of efficiency in producing its outcome. Nevertheless, such actions could also be portrayed as exhibiting rationality by assuming that they are driven by goals that are unattainable in the actor's situation or by beliefs that do not match reality. In other words, by knowing the desire that motivates an action (the chicken wanted to brood her eggs), one can attribute beliefs to the actor that make her seemingly futile action rational (she thought the eggs were on the other side of the road), or by knowing her beliefs (she thought that the other side was free from predators), one can posit a desire (safety) to rationalize her behaviour. In fact, even if neither the actor's desire nor her beliefs are known to us, we can still come up with such inferred mental states to explain her actions. How do we do this?

Writing in *Nature Human Behaviour*, Baker *et al.*⁶ develop a Bayesian computational model to account for the sophisticated inferences we make about the mental states of the actors whose behaviour we observe. The model assumes that actions are generated via a rational planning process

aiming to fulfil the actor's current desire at the lowest possible cost in the world depicted by her beliefs. In this model, inferring the actor's mental states is a kind of inverse planning⁷: finding a set of causal factors (in this case, beliefs and desires) that would produce the observed action as their effect. Normally, such inverse inferences are ill-defined problems that would have too many solutions without further constraints. However, Baker *et al.* account for such constraints by including 'visual perception' in their model: the modelled actor could modulate his beliefs by the segments of the world that he could 'see'. With this approach, the attributed beliefs are determined both by forward inferences from the actor's view of the world ('seeing is believing') and by backward inferences from his observed actions in this world (beliefs that justify the actions). Further, as Baker *et al.* demonstrate in their second study, such a model can be adopted to make inferences not just about the mental states of the actors but also about the world in which he is acting — with the assumption that he acts rationally on the basis of his beliefs, which in turn are causally dependent on his percepts of the real world.

Baker and colleagues' Bayesian model predicts with impressive accuracy the inferences that people would make when observing actions. In this sense, the model can be seen as an approximation of what psychologists call the theory of mind (ToM): a set of computational resources that humans use for attributing representational mental states, such as beliefs and desires, to each other. Does this conclusion imply that the human ToM is Bayesian in nature? Perhaps it does. But note that both the model and the participants in these studies enjoyed the privilege of having to operate over a small set of pre-defined hypotheses concerning the potential contents of the mental states (and concerning the states of the world). It would certainly be easier to interpret the chicken's behaviour if her potential desires were already restricted. Will the author's model simulate human

ToM well if a multitude of potential desires and beliefs compete to explain an observed action? Or are there ways to generate these hypotheses from generic knowledge about the actor's preferences and competencies?

These questions aside, the work by Baker *et al.* opens up promising avenues for testing further questions about the nature of human social cognition. For example, rational planning may sometimes require an actor to pursue sub-goals that may not be attractive in themselves, but are necessary to create conditions that enable some highly beneficial outcome. Some of these sub-goals would not even change the world, but would simply provide information for the actor as to how to optimize her actions ('the chicken crossed the road to get a better view of the terrain'). Indeed, Baker *et al.* discuss the need to consider epistemic goals in future. Furthermore, many human actions are not solitary events, like the ones modelled in this study, but social interactions, which can be costly or beneficial to both the actor and other parties. Whether or not the Bayesian ToM model proposed in this study can be extended to account for such more complex interactions is a question for future research. But the success of this model already demonstrates the fruitfulness of taking seriously the insight that action explanations implicitly assume rational planning. □

Gergely Csibra is in the Department of Cognitive Science, Central European University, 1051 Budapest, Hungary.

e-mail: csibrag@ceu.edu

References

- Davidson, D. *Essays on Actions and Events* (Oxford Univ. Press, 1980).
- Dennett, D. C. *The Intentional Stance* (Bradford Books, 1987).
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. *Trends Cogn. Sci.* **20**, 589–604 (2016).
- Heider, F. & Simmel, M. *Am. J. Psychol.* **57**, 243–259 (1944).
- Liu, S. & Spelke, E. S. *Cognition* **160**, 35–42 (2017).
- Baker, C. L., Jara-Ettinger, J., Saxe, R. & Tenenbaum, J. B. *Nat. Hum. Behav.* **1**, 0064 (2017).
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. *Cognition* **113**, 329–349 (2009).

Competing interests

The author declares no competing interests.