Heroes of our own story: Self image and rationalizing in thought-experiments

Tomer D. Ullman

Abstract: Cushman's rationalization account can be extended to cover another part of his portrayal of representational exchange: thought experiments that lead to conclusions about the self. While Cushman's argument is compelling, a full account of rationalization as adaptive will need to account for the divergence in rationalizing one's actions compared to the actions of others.

Suppose that, like Jason Bourne, you find yourself without memory. In an unfamiliar hotel room, two bodies at your feet, blood splattered on your clothes. In your right hand -- a gun. In your left hand -- a copy of Cushman's "Rationalization is Rational". What should you do now?

Cushman argues persuasively that it is entirely reasonable to go about taking actions in the world, in order to recover a policy from actions driven by non-beliefs and non-desires. You can leave your hotel room and start acting and reconstructing, armed with the knowledge that through evolution and habituation your actions probably make sense (also, armed with a gun). But, you could also sit down on the hotel bed and have a think.

Thought experiments are part of representational exchange, according to Cushman, a way of learning about the world that is "beyond decision making". There are different accounts of how learning from thought experiments works, with many that suggest an explicit unpacking of mental models that have implicit constraints (see e.g. Mach, 1987; Clement, 2009; Lombrozo, in press). However, thought experiments often do involve decision making, and the knowledge we gain through them is not necessarily about the world, but about the self. Here I have in mind thought experiments of the more everyday sort, the 'would you rather' questions that people like to engage in, as opposed to 'what would two blocks tied together to a string do when falling' that only very specific people like to engage in. But many moral reasoning problems fall under this category as well.

Cushman's framework can help explain why such everyday thought experiments are informative, and also why people like to engage in them. Assume that people do not have direct access to their own underlying reasons for action (whether beliefs, desires, habits, or something else) but rather construct a kind of belief-desire theory of themselves (e.g. Gopnik & Meltzoff, 2006; Saxe, 2009). A thought experiment that asks the thought experimenter what action they would take can engage non-rational (habitual, evolutionarily granted) decision making modules, to produce a hypothetical decision. This decision can in turn be used to update a person's theory of themselves, through a similar mechanism to the inversion of the reward from real actions for others (Baker et al., 2009, 2017). But all of this would be happening without setting foot out of the room. In the same way that Cushman posits an 'offline planning' direction from planning to habit in representational exchange, this may be an 'offline rationalization'.

People take pleasure in answering such thought experiments (McCoy, Paul, and Ullman, 2019) because information gained in this way is rewarding, in the same way that any information gain or uncertainty reduction may be rewarding in and of itself (Auer, Cesa-Bianchi & Fischer, 2002). This dynamic of answering from inaccessible modules and updating a theory of those modules can also explain how people can surprise themselves in such thought experiments (McCoy, Paul, and Ullman, 2019), to the degree that there is a misalignment between the two.

So, Bourne could order some room service, pick up a book, and learn something about himself. However, the overall rational-rationalization account as inverse-policy-learning leaves out a possible central constraint that seems different for inverting one's own policy compared to inverting the policy of another: people are the heroes of their own story. When seeing unconscious cops at his feet, Bourne could reasonably conclude he's a bad person. Anyone walking into the room at that moment would likely draw that conclusion, so why doesn't Bourne? Fanciful stories aside, there are many situations in which similar behavior driven by similar habits in ourselves and others are rationalized differently, in a way that is skewed in our favor. When I fail to study for a test, it is because the material was not engaging. When you fail to study, it is because you don't like to work hard. In reality, we were both just tired and hungry. Rationalization-is-rational can explain why people try to reconstruct mental variables in these situations as an adaptive behavior, but to the degree that it is adaptive through being often accurate, it seems odd that rationalization would often diverge in this systematic way. Unless there was some additional difference to make a difference in this computation. And that difference would itself need to be explained on adaptive grounds.

This is a question about what, if anything, needs to go into the inversion of the policy to make rationalization different for myself and others. It is possible that this is a matter of different input information or missing information in that calculation: In addition to the action of not studying, I am also privy to certain mental states like the fact that I am not lazy. But this seems to be begging the question; the whole point of rationalization is that it reconstructs such states where there were none, without the awareness of the person doing the rationalization. An alternative is that there is an overarching, adaptive principle that ensures that rationalization for one's self is more in one's favor than inverse planning for others. This would be akin to a prior on one's own beliefs and desires being in line with what one sees as good or desirable. But if this prior comes at the expense of accurate inference, why have it at all?

In short, I am on board with Cushman's account of rationalization as adaptive, and as part of a broader account of representational exchange. If anything, I think this account can be broadened to capture the engaging aspect of thought experiments that involve making a decision. However, a full account of the functional role of rationalization will need to account not only for its self-benefitting nature, but also its self-serving one.

1. References:

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*(2-3), 235-256.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Clement, J. J. (2009). The role of imagistic simulation in scientific thought experiments. *Topics in Cognitive Science*, *1*(4), 686-710.

Gopnik, A., & Meltzoff, A. N. (2006). Minds, bodies, and persons: Young children's understanding of the self. *Self-awareness in animals and humans: Developmental perspectives*, 166.

Lombrozo, T. (in press). "learning by thinking" in science and everyday life. In *The Scientific Imagination*. Oxford University Press.

Mach, E. [1897/1976]: 'On Thought Experiments', in Knowledge and Error, sixth edition, Dordrecht, Reidel, trans. T. McCormack and P. Foulkes

McCoy, J., Paul, L., & Ullman, T. (2019). Modal prospection. *Metaphysics and Cognitive Science*.

Saxe, R. (2009). The happiness of the fish: Evidence for a common theory of one's own and others' actions. *The handbook of imagination and mental simulation*, 257-266.