Rationalization: Why, when and what for?

Rebecca Saxe¹ and Daniel Nettle²

¹Department of Brain and Cognitive Sciences, MIT

²Institute of Neuroscience, Newcastle University

I can be puzzled by my own action. In this circumstance, I can't directly introspect my reasons, but I can infer plausible beliefs and desires from observing my actions. This is called rationalization. Cushman's central argument is that rationalization has a particular benefit: from rationalizing my actions, I learn true and useful information. The puzzling actions were generated by mechanisms that are not accessible to introspection, but are adaptive, serving my actual interests by increasing my capacity to survive and prosper. Thus through rationalization of my unplanned actions, I could come to make better future plans.

We have three questions about this idea.

When faced with another person's puzzling action, observers only sometimes rationalize it, assuming that the action was an efficient way to achieve the person's goals given her costs (Gershman et al., 2016). On other occasions, observers may explain puzzling behaviour by writing it off as irrational. Indeed, observers too often infer that that people who chose unfamiliar actions don't just have different beliefs and desires, but are fundamentally irrational (Kennedy and Pronin, 2008). Just as we can explain away others' actions as irrational, we can do the same for ourselves; I could interpret my own puzzling action as the result of a habit, or emotional aversion, or descriptive norm ('I always do that'; 'it just feels bad'; 'it's what we do'). So, our first question for Cushman is: can we predict when observers do, and when they don't, rationalize (their own, or others') puzzling actions?

Perhaps we rationalize when the benefits of doing so are the greatest, but when is that? The target article provides surprisingly little empirical evidence of when and how the promised benefits of rationalization actually accrue. Cushman proposes that rationalization lets me learn useful new information, that I can incorporate into better subsequent plans. This idea ought to make testable predictions: training or enhancing rationalization should cause improved planning; limiting or preventing rationalization should impair planning; and people who more frequently rationalize their past actions should therefore make better plans in future. Some phenomena do seem consistent with these predictions. For example, cognitive behavioral therapy, which may be a means of enhancing rationalization, is an effective strategy for reducing self-harm and improving problem solving (Hawton et al., 2016); and people who can give more distinct and differentiated descriptions of their emotions appear to cope better with unexpectedly

negative or threatening events (Kashdan et al., 2010; Pond Jr et al., 2012; Zaki et al., 2013).

In general, we would like Cushman to translate his framework into more specific, falsifiable predictions.

On the other hand, even if some rationalization can generate some benefits for planning, more rationalization probably doesn't always generate more benefits (Trapnell and Campbell, 1999). Too much rationalization is called rumination: persistently chewing the cud of one's past actions is associated with depression and anxiety, not particularly good planning (Nolen-Hoeksema, 2000). In sum, our second question for Cushman is: under what circumstances do people who rationalize their actions actually make better plans? And how should we measure 'better plans', to test this idea?

There may be a link here to the concept of 'resource rationality' (Griffiths et al., 2015). The principle of resource rationality is to achieve the best cognitive outcome, counting the computational costs. A resource-rational creature would thus rationalize specifically when the benefits outweigh costs of more extensive computation; for example, where situations have complex and contingent payoff structures; where they involve close relationships of high value; around highly consequential decisions made under uncertainty; in the wake of conflict with significant others; or when following instincts, learned associations or norms leads to a large error signal. It would be valuable to understand empirically if these are the cases where most rationalization does indeed occur. The idea of resource rationality also links quite naturally to the possibility of pathological over-rationalization: this would occur when rationalization resources are deployed beyond the point where their marginal benefit for future situations could exceed their cognitive cost.

Finally, we feel the target article may underestimate the importance of the interpersonal functions of rationalization, as well as how tight the linkage is between inter- and intra-personal ones. Cushman considers one way rationalization operates in interpersonal interaction: when I offer an exculpatory but false explanation of harmful actions. In this context, since I know my explanation is false, the only reason I might actually acquire the rationalized beliefs and desires is to achieve more effective deception of others (Trivers, 2000). We suspect that this example is too narrow, missing much of the proximal motive and the ultimate benefit of rationalization.

Like Cushman, we expect that human minds contain multiple mechanisms vying for control of behaviour. In our view, such competition occurs not just between separate modules with different functions. There can also be conflict within any one system, for example between competing desires, fears, norms and expectations. Inner conflicts arise because people are uncertain about the present, about the future, about the value of key rewards, and so on.

The benefits of rationalizing one's own actions could arise mostly from simplifying the sources of our actions that we share with other people. When a person's actions actually emerge from inner conflict and competition, the resulting sequence of behaviours could be quite confusing to third-party observers trying to infer a unitary set of beliefs and desires. Rendering our own actions seemingly rational and comprehensible could have many interpersonal benefits, including signalling our value as a future partner, and persuading others to subscribe to joint plans (Mercier and Sperber, 2011).

In short, the target article stimulates further questions: about when rationalization is

most likely to occur, and to not occur, and about where to expect, and how to measure, its benefits.

References

- Gershman, S. J., Gerstenberg, T., Baker, C. L., and Cushman, F. A. (2016). Plans, habits, and theory of mind. *PLoS one*, 11(9):e0162246.
- Griffiths, T. L., Lieder, F., and Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2):217–229.
- Hawton, K., Witt, K. G., Salisbury, T. L. T., Arensman, E., Gunnell, D., Hazell, P., Townsend, E., and van Heeringen, K. (2016). Psychosocial interventions following self-harm in adults: a systematic review and meta-analysis. *The Lancet Psychiatry*, 3(8):740–750.
- Kashdan, T. B., Ferrisizidis, P., Collins, R. L., and Muraven, M. (2010). Emotion differentiation as resilience against excessive alcohol use: An ecological momentary assessment in underage social drinkers. *Psychological Science*, 21(9):1341–1347.
- Kennedy, K. A. and Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, 34(6):833–848.
- Mercier, H. and Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory.
- Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of abnormal psychology*, 109(3):504.
- Pond Jr, R. S., Kashdan, T. B., DeWall, C. N., Savostyanova, A., Lambert, N. M., and Fincham, F. D. (2012). Emotion differentiation moderates aggressive tendencies in angry people: A daily diary analysis. *Emotion*, 12(2):326.
- Trapnell, P. D. and Campbell, J. D. (1999). Private self-consciousness and the five-factor model of personality: distinguishing rumination from reflection. *Journal of personality and social psychology*, 76(2):284.
- Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, 907(1):114–131.
- Zaki, L. F., Coifman, K. G., Rafaeli, E., Berenson, K. R., and Downey, G. (2013). Emotion differentiation as a protective factor against nonsuicidal self-injury in borderline personality disorder. *Behavior therapy*, 44(3):529–540.