Rationality as the end of thought

Nick Chater

Behavioural Science Group

Warwick Business School

University of Warwick, Coventry CV4 7AL, United Kingdom

nick.chater@wbs.ac.uk

Word counts: abstract: 65, main text, 1035; references, 100; total, 1200).

Abstract

Bermúdez convincingly argues that framing effects are ubiquitous and that this is not a sign of human irrationality, but an unavoidable feature of any intelligent system. The commentary adds that framing effects arise even in formal domains, such as chess and mathematics, which appear paradigms of rational thought. Indeed, finding and attempting to resolve clashes between different frames is a major impetus for deliberative cognition.

Bermúdez makes a compelling case that the impact of framing on reasoning and choice is both widespread and entirely reasonable. Yet finding that one is subject to a framing effect does imply that one is, in some sense, in cognitive disequilibrium: some further thought is required to determine what to think or do. I suggest that theories of rationality, both formal and informal, should rightly be construed as providing conditions for equilibrium (Chater & Oaksford, 2012). So, for example, probability theory, logic, and game theory are all attempts to establishing when our potentially divergent intuitions, prompted by different frames, can simultaneously be embraced. Conversely, these conditions determine when our thoughts are our of equilibrium and adjustments are required (these theories do not, crucially, specify which of the many possible adjustments should be made).

Bermúdez highlights how framing effects arise naturally in the political and ethical dilemmas of literature and real life, as well as being a staple of laboratory experimentation. But it is worth stressing that framing effects are likely arise in any situation in which boundedly rational agent faces a problem that is too complex to be solved completely.

Let us take the example of chess, where the objective and rules of the game are formally specified, and there are widely agreed standards of what counts as a good move (these days, a good move is operationalized by referring to the "chess engines" that now spectacularly outperform human players). During play, a human (or machine) chess player will continually generate and evaluate conflicting arguments for the virtues of different possible moves---and framing effects will be legion. Suppose, for example, a player is considering an innocuous move (say, advancing a pawn). As different possible consequences of the move are considered (i.e., continually shifting and elaborate its framing), the overall evaluation of its virtues may ebb and flow. If the different frames give wildly different answers, the player may either abandon the move as too risky, or think further to establish which frame should dominate. For example, if the frame is "gain control of the center of the board" the move might seem uninspired but solid; but under the framing "trigger an exchange of pawns and then knights, weakening the defence of the opponent's king," it may seem more attractive. Suppose the player decides to make the move, and then is confronted with a completely unexpected queen sacrifice, which leads to checkmate in three moves. Now the earlier move is seen through a different frame, which was not previously considered---and the dismayed player will realize that this frame is decisive.

Is the player displaying irrationality? It might appear so, from the point of view of an extensional decision theory. After all, the player evaluates "advance pawn" as a good move; and moments later "advance pawn, opening up the possibility of a devasting Queen sacrifice" as a bad move. But these are, of course, the same move, simply described differently. But it would entirely misguided to criticize a person for such a mistake, saying: "don't worry about the description, just make the best move," because we can only evaluate whether or not a move is good or not by considering specific descriptions (including strategic advantages, likely countermoves, etc). Indeed, a purely extensional approach to playing chess would entirely "abstract away" from the computational challenge of chess---and the reason that chess requires hard thought in the first place.

What is the role of the game theory here (i.e., as providing a rational theory of how strategic interactions should be played)? I suggest that we should view it as providing no more than mild consistency constraints. For example, if moving the pawn leads to certain defeat (after the

unexpected queen sacrifice), then reasoning backwards, this must have been a bad move, on the assumption that the opponent will choose their move to maximize the chance of winning (and assuming perfect rationality). Similarly, the goodness of the current position should relate directly to the goodness of the position after each player has made the "best" next move; and so on.

But these are minimal constraints ignore almost everything of interest in the game. Indeed, a pure game-theoretic analysis of chess would simply advise each player to choose a winning strategy from the outset (if White or Black has a winning strategy, which is not known); or otherwise each should play out a certain draw (Schwalbe & Walker, 2001). But the computational complexity of chess is such that no such strategies can be found for either player (Storer, 1983).

The same picture arises across domains. We have, inevitably, a plethora of inconsistent mathematical intuitions (Lakatos, 1976), and framing will be crucial. One moment we might consider a theorem fairly plausible (perhaps by analogy with some similar theorem), but when re-framed as having the consequence that Fermat's Last Theorem is false, the credibly of the theorem reduces sharply. The purpose of mathematical reasoning is surely to help uncover and resolve such cases; and progress in mathematics will involve continually generating and resolving inconsistencies, without obvious limit. Similarly, in less purely formal domains, we can see scientific theories, ethical principles, and indeed the project of philosophy itself, as attempting to find and resolve the endless clashes between our diverging intuitions.

A mind without framing effects would be in perfect equilibrium. Principles of rationality would therefore be satisfied for such a mind. But if rationality constraints were fully satisfied, the need for further thought would have come to an end. In any case, such equilibrium is unattainable. Framing effects, and our continual and partial attempts to resolve them, are not a signature of irrationality; rather they are inevitable consequence of grappling with a world more complex than we can fully understand (cf. Harman, 1986).

Rationality is also the end of thought in a rather different, and more positive sense: the objective of reconciling different frames to be rationally consistent is a driving force behind deliberative cognition. Searching for and evaluating different frames demands intense thought by chess players or mathematicians, and lengthy soliloquizing by Agamemnon and Macbeth. Indeed, it is no exaggeration, perhaps, to see the resolution of conflicts between frames as a major driver of individual and collective cognitive progress.

Conflicts of interest: None

Funding: This works was supported by the ESRC Network for Integrated Behavioural Science [grant number ES/K002201/1].

Chater, N. & Oaksford, M. (2012). Normative systems: logic, probability, and rational choice. In K. Holyoak & R. Morrison (Eds.). *The Oxford Handbook of Thinking and Reasoning* (pp. 11-21). New York: Oxford University Press.

Harman, G. (1986). *Change in view*: *Principles of Reasoning*. Cambridge, MA: MIT Press.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge, UK: Cambridge University Press.

Schwalbe, U., & Walker, P. (2001). Zermelo and the early history of game theory. *Games and Economic Behavior*, *34*(1), 123-137.

Storer, J. A. (1983). On the complexity of chess. *Journal of Computer and System Sciences*, *27*(1), 77-100.