



CHAPTER

33

c00033

Remarks on Neuroeconomics

Daniel Kahneman

OUTLINE

s0010

Acknowledgments

524

References

524



p0010

I have been an enthusiastic and ignorant fan, since the early days, of what used to be called cognitive neuroscience, and an even more ardent fan since an event that happened, so it seems, yesterday – the skyrocketing takeoff of neuroeconomics. Over the years I have had many difficult conversations with colleagues in my discipline, who share my ignorance but not my enthusiasm. The skeptics are quick to agree that thinking and choice are outcomes of events that occur in the brain; they question the usefulness *to psychology* of knowing which parts of the brain are activated when particular thoughts or choices happen. (The argument is strikingly similar to the case that some economists have made in favor of a “mindless economics”: psychological data should be ignored because economics is a self-contained discipline (Gul and Pesendorfer, 2005). A case for “brainless economics” should be even easier to make.)

p0020

Imaging studies provide the main source of data for neuroeconomics, although there is hope that this could change as new techniques are introduced. To be useful to psychology, of course, such measures of brain activity must be translated into psychological

terms. One of the happy surprises of neuroeconomics is the frequent finding of impressive correlations between psychological measures and measures of brain activity (or, sometimes, of differences between levels of activity in different regions). Among other examples cited in this book are a within-subject correlation between the rejection of unfair offers and the activation of right anterior insula (Sanfey and Dorris), a high correlation between a behavioral measure of loss aversion and a neural measure of loss aversion in ventral striatum and ventrolateral prefrontal cortex (Fox and Poldrack), and a correlation between the *difference* in neural measures of people’s utility for their own money and for others’ money and a willingness to donate to these others (see Chapter 20 of this volume). The within-subject correlations suggest an impressive ability to identify a close correlate of a causal mechanism.

The between-subject correlations help to make another point, on a topic on which psychologists and economists have very different perspectives: the question of interpersonal comparisons of utility. If brains are sufficiently similar in structure and function to



support high correlations between indices of brain activity and measures of psychological states, the calibration function that relates the psychological to the neural variables is unlikely to be very different across individuals, and there appears to be little justification for the taboo regarding interpersonal comparisons.

p0040 High correlations between well-identified psychological and neural measures are the exception, not the norm. In most experiments, of course, the correspondence between psychological terms and neural measures is more equivocal, and the interpretation of imaging results is tricky. Russ Poldrack (see Chapter 11 of this volume; also Poldrack, 2006) has drawn attention to the problem of “reverse inference,” which arises when people infer a specific psychological process from activity in a particular region – for example, when activity in dorsal striatum is interpreted as an indication that people enjoy punishing strangers who have behaved unfairly (De Quervain *et al.*, 2004; see also Chapter 17 of this volume). There is indeed a problem, because activity in ventral striatum is not perfectly correlated with enjoyment: many other circumstances produce activity in that region, and there is no assurance that it will be active whenever the individual experiences pleasure. In spite of this difficulty, the result and its proposed interpretation are just what a general psychologist (not a neuroscience specialist) would order. It is surprising but plausible, and it drives thinking in new directions. The more difficult test, for a general psychologist, is to remember that the new idea is still a hypothesis which has passed only a rather low standard of proof. I know the test is difficult, because I fail it: I believe the interpretation, and do not label it with an asterisk when I think about it. And I will be sorry if it is disproved, but will have no difficulty in accepting its demise – it would join a long list of defunct once-cherished ideas.

p0050 In spite of the fact that many regions of the brain respond promiscuously to many experimental situations, the interpretation of positive results seems relatively tractable. I like the analogy of a personality inventory, a collection of hundreds of questions with dichotomous answers (my memory from graduate school is that the item “I used to enjoy ripping the wings off flies” (or words to that effect), when answered in the negative is a weak indication of depression). Every brain region is like an individual, answering “yes” or “no” to thousands of experimental questions that are devised by the scientific community. The *pattern* of yes–no answers – where significant activation is “yes” – is an indication of the “personality” of a region. As the evidence accumulates, it seems very likely that practitioners of the art will develop an

intuitive appreciation of these personalities – a sense of how a region is likely to respond to a new situation – which should help in the critical stage of selecting regions of interest. It is also likely that computational techniques will be applied in an attempt to organize the explosively growing amount of positive findings and ideas.

The interpretation of negative results presents a more difficult challenge. The asymmetry between the treatment of positive and negative findings is a major feature of the current situation, probably unavoidable in the early days of a developing methodology. However, it is a problem, as I well remember from my only foray into the field (Breiter *et al.*, 2001). Loss aversion is a salient aspect of the behavioral analysis of decisions, and I had expected to see manifestations of loss aversion showing up in vivid colors on the images. I had also anticipated that outcomes would be more salient than expectations – which they are, of course, in our subjective experience of life. None of this happened. Gains showed up more than losses, and the sight of a favorable or unfavorable prospect caused more activation than its eventual resolution. These were clearly important observations, but I did not know what we could make of them that would change psychological thinking about decision-making. The similarity to the reactions of the dopamine system (see Chapter 21) provided a context for the relative salience of predictions, although it did not resolve the apparent inconsistency with subjective experience. The absence of obvious manifestations of the powerful psychological response to losses has remained a problem, although several chapters in the book report advances (see Chapters 7, 11, and 12; see also Knutson *et al.*, 2007).

Early in the days of any field or any technique, the most valuable results are those that confirm expectations or help to sharpen existing hypotheses. As the field and its techniques mature, of course, believable surprises become the goal and the prize. My impression is that, at least in the domains with which I am familiar, it is still early days for neuroeconomics. When I was asked by a well-known critic of the field what I had learned that had changed my mind about decision-making, I did not have much to say. In part, of course, it is because I was right all along! The findings of neuroeconomics research have generally confirmed the expectations of behavioral decision theorists and behavioral economics. However, we are beginning to learn more, and I am confident that the pace will accelerate in coming years. One example of what we are learning concerns the interpretation of framing effects. When people are presented with the

same choice between a sure thing and a gamble, they prefer the sure thing if the outcomes are framed as gains and they prefer the gamble if the outcomes are phrased as losses. Why? Some recent findings and theoretical analyses suggest an interpretation: a primary tendency to hate sure losses and to be attracted to sure gains, when such simple outcomes are compared to more complicated risky prospects. In this volume, the idea that these preferences are due to a simple Pavlovian response is suggested by both Dayan and Seymour (Chapter 13), and Bossaerts, Preuschoff, and Hsu (Chapter 23). An imaging study of framing effects (De Martino *et al.*, 2006) led Shane Frederick and myself to a similar interpretation (Kahneman and Frederick, 2007), which gains further indirect support from the ingenious studies of framing effects in capuchin monkeys that Santos and Chen report in Chapter 7 of this volume. Another topic on which neural data may help to articulate theory concerns the rewards that guide actions. This is one of the main issues of contention between behavioral economics and the standard economic model. The notion that preferences are constructed in the immediate context stands in sharp contrast to the idea of a preference order which is basic to rational-choice theories. The concept of actions that are their own rewards – altruistic punishment and charitable giving may be examples – is, of course, much more compatible with the former than with the latter, and there is growing support for it in neural data (see Chapters 6, 15, and 20). The very meaning of rationality changes when actions are determined by the intrinsic utility they provide. I believe we are well on our way to the day where our theoretical concepts about decision-making are shaped at least in part by findings from neuroscience.

p0080

The difficulties associated with the conceptual interpretation of neuroscience evidence can be traced to the correlational nature of most of that evidence. A new era in neuroeconomics began with the introduction of experimental manipulations such as transcranial direct current stimulation (tDCS) or the inhalation of oxytocin to induce predictable behavioral consequences (see Chapter 15). Fehr's justifiable enthusiasm is captured in this quotation:

A key feature of tDCS is that it is inexpensive and can be simultaneously applied to many subjects who interact in a laboratory environment (Knoch *et al.*, 2007). Thus, in principle, tDCS can be applied to a group of, say, 20 subjects simultaneously, with each of them playing one one-shot game with the other 19 subjects. Therefore, tDCS could prove to be a non-invasive brain stimulation method that revolutionizes neuroeconomics because it greatly enhances data-collection efficiency and enables brain stimulations in whole groups of interacting subjects.

The application of minimally invasive interventions that alter brain function will drastically increase the relevance of studies of the brain to an understanding of the deciding mind. The only development that could do as much for the field would be a qualitative improvement in the temporal resolution for patterns of activation. p0090

The instructive chapter that introduced this volume mentioned two studies published in 2001 (Breiter *et al.*, 2001; McCabe *et al.*, 2001) that announced (they did not cause) two strands of research which are prominent in this collection: studies of social preference in contexts defined by game theory, and studies of valuation of uncertain outcomes. To say that a great deal has happened since then is an obvious understatement, but the progress has not been even. The valuation and decision problem is broader and harder, and a meaningful theoretical integration seems much more remote. For one thing, there is no agreement on whether the theoretical language for such an integration will come from computational neuroscience, from rational choice theory, or from behavioral economics. In Chapter 32, Paul Glimcher offers utility theory as a guiding framework for the field, while acknowledging that there is much it does not cover. His proposal will certainly be contested. A unitary concept of utility cannot do justice to the complexities of the relationships between what people (and other animals) want, what they expect to enjoy, what they actually enjoy, and what they remember having enjoyed. Unless one believes that there is never a difference in these types of utility for anybody's decision, neuroeconomics might provide some of the best tools to learn when they differ and why. Furthermore, the hypothesis of optimal performance is not equally applicable to all domains. We know, for example, that the human perceptual system is more reliably Bayesian than is human judgment, and optimality in the organization of movement (see Chapter 8) or in the automatic allocation of attention does not necessarily generalize to other categories of performance. p0100

What do we do when behavioral evidence and neural evidence point in different directions? Some time in the future, I expect, such discrepancies may come to be resolved by reinterpreting the behavioral data to fit the conclusions from the neuroscientific work. This day has not come yet. Here again, in Chapter 32, Glimcher raises an interesting challenge. He believes that there is no evidence from neuroscience to support the idea that (some) decisions arise from a conflict between emotion and reason, and concludes that a unitary system will do the job. His position on the neuroscience is not universally held (see, for

example, McClure *et al.*, 2007), but this is not my point. I start from the position that there is overwhelming behavioral evidence both for the existence of multiple systems of thought and choice and for the importance of conflict (Kahneman, 2003). As of today, the absence of repeated studies showing well-localized neuroscientific correlates for multiple systems should still be viewed as an unsolved problem for neuroeconomics, rather than as a problem for the idea of multiple systems. I emphasize “as of today,” because I firmly believe that the findings of neuroscience – negative as well as positive – will soon play a large role in shaping the concepts and theories of behavioral research.

s0010 Acknowledgments

p0120 I thank Colin Camerer, Craig Fox and Russ Poldrack for helpful comments and suggestions. The usual caveats apply.

References

- Breiter, H.C., Aharon, I., Kahneman, D. *et al.* (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30, 619–639.
- De Martino, B., Kumaran, D., Seymour, B., and Dolan, R.J. (2006). Frames, biases, and rational decision-making in the human brain. *Science* 313, 684–687.
- De Quervain, D.J.-F., Fischbacher, U., Treyer, V. *et al.* (2004). The neural basis of altruistic punishment. *Science* 305, 1254–1258.
- Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev.* 93, 1449–1475.
- Kahneman, D. and Frederick, S. (2007). Frames and brains: elicitation and control of response tendencies. *Trends Cogn. Sci.* 11, 45–46.
- Knoch, D., Nitsche, M.A., Fischbacher, U. *et al.* (2007). Studying the neurobiology of social interaction with transcranial direct current stimulation: the example of punishing unfairness. *Cerebral Cortex Advance Access*.
- Knutson, B., Rick, S., Wimmer, G.E. *et al.* (2007). Neural predictors of purchases. *Neuron* 53, 147–156.
- McCabe, K., Houser, D., Ryan, L. *et al.* (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl Acad. Sci. USA* 98, 11832–11835.
- Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* 10, 59–63.