Avril Kenney

10/21/08
Paper 2

How You Think You Think:
The Neuroscience of Irrationality and Introspection

Introduction

Recent studies in psychology, economics, and neuroscience have provided substantial evidence that people's decision-making process is not rational. Such findings challenge many classic theories of human behavior, but there is little question about their validity. This disconcertingly large amount of irrationality is forcing economists, cognitive scientists, legal policymakers, and many others to reassess how they have been modeling people's behavior.

Imagine, however, a foreign being observing humanity. To him, the surprising part of these events would not be the discovery that we are irrational – that much would have been obvious. What would surprise him is that we envision ourselves to be otherwise. The very fact that these findings are new and shocking, that they draw attention and spark controversy, that they are in opposition to assumptions on which we have based our society, indicates that we do not know nearly as much about ourselves as we think we do.

In fact, current neuroscientific research is suggesting that we do not directly understand ourselves at all. Just as we have expectations for what a ball does when it is dropped, we have expectations for what a person does when he is faced with a choice. These expectations are based on our mental representations of the entities involved. In the case of the ball, we cannot see the force of gravity, but we infer its existence from observing its effects. In the case of the person, we cannot see the computations taking place in his mind, but we use logical reasoning to come up with feasible explanations. Taking it a step further, the same argument can be applied to oneself: we cannot see the causes of our decisions, but we know the circumstances and the result, so we make up reasons consistent with the information we have.

Not surprisingly, this idea – that people are only sometimes correct about how they themselves reach their decisions – makes many people uncomfortable. It raises philosophical questions about the nature of free will and responsibility, about what constitutes the "self." After all, if you cannot even explain why you did something, how can you be held accountable for a decision that your brain made without you? In reality, though, such concerns are unwarranted. A person's brain is a part of himself, and the fact that the decision-making process is below the level of consciousness does not exempt him from responsibility for its results.

Irrational Decisions

Humans possess the powerful ability to use rule-based logic to determine new facts based on known information. It has therefore traditionally been assumed that such logical methods are how we reach all of our conclusions. Fields that have to anticipate general patterns of how humans will behave, such as economics and political science, have used models in which people simply take into account the facts they have and follow them to the logical conclusions. In these models, information is represented in an objective manner and is ignored when irrelevant. Logically, people should assess value on a constant scale and decide on the outcome that maximizes value. However, studies of human behavior show that this is not how we actually think.

In stochastic situations, where the outcome is uncertain, one can calculate the "expected value" of the result, which is what one would get on average. When given a choice among multiple events, a person should choose the one with the highest expected value – but this is not what happens. People have a strong desire to avoid to negative consequences, so they tend to select options with lower probabilities of very bad outcomes, even when the overall expected value is worse. Similarly, they tend to prefer somewhat good outcomes that have high certainty over much better outcomes that have lower certainty. Rangel, Camerer, and Montague (2008)

describe two models, Expected Utility Theory and Prospect Theory, that have been proposed to explain these observations. Both appear to have neural correlates in the human brain, implying that we have a dedicated system for reaching decisions of this type.

Another factor influencing people's choices is distance in time. When given the option of receiving $10 today or $11 tomorrow, many people would choose the slightly smaller amount at the sooner time; but if the times were 365 days from now versus 366 days from now, the same people would likely choose the larger amount of money, even though the time difference is the same. McClure et al. (2004) used fMRI to study people's brain activity when making such decisions, and found that the immediate rewards tended to activate areas in the limbic system, while delayed rewards activated more cortical regions.

The time-discounting example shows that value is often measured relative to some baseline as opposed to on an absolute scale – relative to a year, one day does not make much of a difference – and in fact, our baseline values can be quite manipulable. Ariely, Loewenstein, and Prelec (2006) performed a study in which subjects had to write down the last two digits of their social security numbers, decide whether they would pay that amount in dollars for a specified item, and then state the maximum amount they would pay for the item. Despite the fact that social security numbers have no relation to the question, people with higher numbers were on average willing to pay much larger amounts, indicating that our judgments can be "anchored" at arbitrary starting points.

It is not only in valuing physical items or monetary amounts that our decisions disobey the rules of logic: much of what we call morality is highly irrational as well. In the classic "trolley problem," people are asked whether they would pull a lever to make a trolley hit one person instead of five, and whether they would push a person in front of the trolley to prevent it from hitting five others. Most people answer yes to the first question and no to the second, even

though the two actions would have the same outcome. Greene et al. (2001) propose that this difference is caused by emotions elicited in the subject by the stories, and show that different brain areas are activated during decision-making in emotional versus non-emotional moral dilemmas.

Rational Explanations

Clearly, under many circumstances people's decisions are not rational. Why is it, then, that for such a long time we based entire areas of study on the assumption that they are? If we could directly observe the underlying causes of our actions, we would immediately realize the invalidity of such an assumption. Thus, it follows that people are not automatically capable of knowing why they do what they do, even within themselves.

In the example of the trolley problem described above, many people are unable to provide any reason for why they chose their answers. Even when they adamantly believe in their choices, they cannot give any logical explanation, simply because one does not exist. Hauser et al. (2007) performed a study in which subjects were asked to judge the permissibility of killing a person in variants of the trolley problem, and then to provide justifications for their decisions. Of the subjects who gave differing answers to the two versions (excluding those who added extra assumptions not stated in the question), only 30 percent were able to give adequate justification for their judgments. Furthermore, subjects who had exposure to readings on moral philosophy were significantly more likely to sufficiently justify their positions than those who had not, even though the two groups had the same likelihood of judging the acts permissible. This implies that people use external rather than internal information to come up with explanations for their decisions.

Additional compelling evidence comes from Nisbett and Wilson (1977), who describe a comprehensive set of studies showing that people often give incorrect accounts of what factors

influence their decisions. As just one of many examples, in one experiment subjects were shown four pairs of stockings and had to decide which was the best quality and why they thought so. The pair farthest to the right was about four times as likely to be chosen as the one on the left (the stockings were all objectively of identical quality), yet none of the subjects cited position as having played a role in their decision-making process, and almost all of them denied it even when specifically asked.

A more recent study by Johansson et al. (2005) shows that people give qualitatively equivalent explanations for their choices regardless of whether the choices were ones they really made. Subjects were shown two photos of faces and asked which was more attractive; they were then given the photo they had selected and asked to explain why they had picked it over the other. But in some of the trials, the experimenter switched the photos, giving the subject the one he had not actually chosen. People usually did not detect the switch, and there was no significant difference between their explanations for the decisions they had made and the decisions they had not made (measured by ratings of emotionality, specificity, and certainty). This finding implies that people have no direct evidence of where their decisions come from and are employing the same methods on their own mental processes as they would on any other observed events.

Neuroscientific Evidence

With new imaging technology, it is becoming possible to localize even very specific mental functions to separate areas of the brain. As described above, several regions have already been identified as pertaining to different types of decision-making. Now, research is being done to find out what goes on in our brains when we make decisions versus when we think about them.

Soon et al. (2008) conducted an fMRI study in which subjects watched letters flash on a screen and decided when to press one of two buttons. For each time a subject pressed a button,

he noted which letter had been on the screen when he became consciously aware of the decision. Two cortical brain regions – one in the frontopolar cortex and one in the parietal cortex – were found to reliably predict which of the buttons would be pressed, up to 10 seconds before the choice entered the subject's consciousness. Given such a delay, it is apparent that people do not directly see the procedures by which their decisions are made, since the information is present in the brain before they are aware of it.

Research has established that certain parts of the brain – particularly the right temporo-parietal junction (right TPJ), but also the left TPJ and medial prefrontal cortex – are selectively recruited when thinking about people's thoughts as opposed to their actions, characteristics, or physical states. Importantly, these areas are distinct from those involved in planning and performing actions, as well as from the mirror neuron system. Although the studies have primarily involved thinking about other people, a few experiments reported by Saxe (in press) suggest that these brain regions might correspond to thinking about mental states in general, whether one's own or someone else's. As of yet there have not been imaging studies requiring subjects to reflect on their own decision-making, so it remains to be seen whether this hypothesis holds.

Philosophical and Moral Implications

Behavioral studies have shown that people's decisions do not arise from a rational process, and imaging experiments are revealing complex and varied neural systems responsible for what we actually do. At the very least, we know that people make irrational decisions and that they cannot always use introspection to figure out why. In light of recent findings, it is probable that we only speculate and do not *ever* know how we come to our decisions, and possible that analysis of one's own behavior is only more accurate than analysis of someone else's to the extent that one has more available external information. The key point is that a

person's mental representation of himself is separate from, and significantly less rational than, his actual self.

It is interesting to note that even after gaining this new insight on how their thinking works, people persist in viewing themselves as free-acting, rational agents. Perhaps it is *because* of our lack of access to causality information that our models of ourselves are so non-deterministic: a person chooses one out of several outcomes, but based purely on reason his choice was not the only one possible, so it must have been that he was *able* to pick any of them and simply *decided* on one. The notion of free will corresponds with this way of experiencing the world. We cannot always predict the future based on the causal laws we know, and we assume that we know all the causal laws shaping our own decisions, so we conclude that our decisions are not always determined by causal laws. Because people are confident that they possess the ability to freely choose among alternatives, they are uncomfortable with the idea that they do not consciously control the outcomes of such choices.

But this disconnect arises from a faulty assumption, namely, that we know the causes of our actions. Once we refute this claim, as the ample research described above has done, free will is a nonissue, for it becomes apparent that so-called free will is merely a nebulous concept we have created to fill the gap between the rational decisions we imagine ourselves to make and the irrational ones we actually make. Our logical minds would not allow us to believe that something can happen with no cause whatsoever, yet often people's actions seem to do precisely that, so we explain them by saying that they are performed of someone's free will.

If everyone had free will as we conceive of it, if everyone's mental representation of himself coincided perfectly with his real-life behavior, it would be easy to hold people responsible for their actions. One could argue that people always have the option of deciding something else, and could assess the reasonableness of a person's decision by having him explain

why he made it. Under the current framework, such an argument no longer holds, since it is not at all clear whether he could have made a different decision, and the rationality of any explanation is more a result of reasoning ability than an account of what really took place. However, this does not by any means exempt people from responsibility; in fact, it supports the idea that all types of people should be held equally responsible for their behavior.

When judging moral responsibility, people have a tendency to differentiate between situations based on external influences, but these influences are often considered to include anything outside conscious awareness – that is, outside the domain of free will. As science identifies a biological basis for more and more psychological phenomena, such as addiction, depression, hyperactivity, and countless others, it is taken as evidence that these are beyond our control, and therefore that we are not responsible for their effects on our actions. In actuality, though, the causes of all our decisions are beyond the reach of our conscious processes, so we cannot use that as a criterion for excluding some actions and not others (or the actions of some people and not others) from responsibility.

Conclusion

People conceive of themselves as consciously aware and in control of their decisions, making logical choices based on rationality and free will. Both of these beliefs are contradicted by robust data from cognitive science, and neuroscience is narrowing in on understanding how the processes occur in our brains. The fact that proposed rational explanations cannot account for our decisions suggests that we should not rely on introspection to deduce the causes of what we do and that rationality should not be the only criterion for judging it. In situations where human behavior has to be predicted or explained, we should consider all that we actually know instead of just what our mental heuristics tell us.

References

Ariely, D., Loewenstein, G., Prelec, D. (2006). "Tom Sawyer and the construction of value." *Journal of Economic Behavior & Organization* 60(1), 1-10.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment." *Science* 293, 2105-2108.

Hauser, M., Cushman, F., Young, L., Jin, R.K., Mikhail, J. (2007). "A Dissociation Between Moral Judgments and Justifications." *Mind & Language* 22(1), 1-21.

Johansson, P., Hall, L., Sikström, S., Olsson, A. (2005). "Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task." *Science* 310, 116-119.

McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D. (2004). "Separate Neural Systems Value Immediate and Delayed Monetary Rewards." *Science* 306, 503-507.

Nisbett, R.E., Wilson, T.D. (1977). "Telling More Than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84(3), 231-259.

Rangel, A., Camerer, C., Montague, P.R. (2008). "A framework for studying the neurobiology of value-based decision making." *Neuroscience* 9, 1-12.

Saxe, R. (in press). "The happiness of the fish: Evidence for a common theory of one's own and others' actions." In K. Markman, B. Klein, J. Suhr (eds.), *The Handbook of Imagination and Mental Simulation*.

Soon, C.S., Brass, M., Heinze, H., Haynes, J. (2008). "Unconscious determinants of free decisions in the human brain." *Nature Neuroscience* 11, 543-545.