Thinking & Reasoning, 2013
Vol. 19, No. 1, 1–26, http://dx.doi.org/10.1080/13546783.2012.713178

Routledge
Taylor & Francis Group

Why humans are (sometimes) less rational than other animals: Cognitive complexity and the axioms of rational choice

Keith E. Stanovich

Department of Human Development and Applied Psychology, University of Toronto, Toronto, ON, Canada

Several formal analyses in decision theory have shown that if people's preferences follow certain logical patterns (the so-called axioms of rational choice) then they are behaving as if they are maximising utility. However, numerous studies in the decision-making literature have indicated that humans often violate the axioms of rational choice. Additionally, studies of nonhuman animals indicate that they are largely rational in an axiomatic sense. It is important to understand why the finding that humans are less rational than other animals is not paradoxical. This paper discusses three reasons why the principles of rational choice are actually *easier* to follow when the cognitive architecture of the organism is simpler: contextual complexity, symbolic complexity, and the strong evaluator struggle.

Keywords: Cognitive complexity; Axioms of rational choice; Decision making; Context and values.

Searle (2001) begins a book on the philosophy of rationality by referring to the famous chimpanzees on the island of Tenerife studied by Wolfgang Kohler (1927). Several of the feats of problem solving that the chimps displayed have become classics and are often discussed in psychology textbooks. In one situation a chimp was presented with a box, a stick, and a bunch of bananas high out of reach. The chimp figured out that he should

Correspondence should be addressed to Keith E. Stanovich, Department of Human Development and Applied Psychology, University of Toronto, 252 Bloor St. West, Toronto, Ontario, Canada M5S 1V6. E-mail: keith.stanovich@utoronto.ca

Preparation of this manuscript was supported by the Canada Research Chairs program and the Grawemeyer Award in Education.

position the box under the bananas, climb up on it, and use the stick to bring down the bananas. Searle (2001) asks us to appreciate how the chimp's behaviour fulfilled all of the criteria of instrumental rationality—the chimp used efficient means to achieve its ends. The desire of obtaining the bananas was satisfied by taking the appropriate action.

Searle (2001) uses the instrumental rationality of Kohler's chimp to argue that, under what he calls the Classical Model of rationality, human rationality is just an extension of chimpanzee rationality. The Classical Model of rationality that Searle portrays is somewhat of a caricature in light of recent work in cognitive science and decision theory, but my purpose is not to argue that here. Instead I wish to stress the point on which Searle and I agree (but for somewhat different reasons). Human rationality is not simply an extension of chimpanzee rationality. However, I do not at all mean to argue that human rationality exceeds that of the chimpanzee. To the contrary, the reverse can often be the case. Nonhuman animal rationality can easily exceed that of human rationality—and in this essay I will discuss why this not at all paradoxical.

There are three reasons why human choices, compared to those of lower animals, might display less of the coherence and stability that define instrumental rationality as it is operationalised in axiomatic utility theory. The reasons are: contextual complexity, symbolic complexity, and the strong evaluator struggle. I will discuss each in turn. The first—contextual complexity—raises the possibility that, due to coding more contextual features into their options, humans risk exhibiting more moment-to-moment inconsistency (of the type that leads to violations of rationality axioms) than less cognitively complex animals whose cognitive systems respond more rigidly to stimulus triggers.

WHY RATS, PIGEONS, AND CHIMPS CAN SOMETIMES BE MORE RATIONAL THAN HUMANS

The model of rational judgement used by decision scientists is one in which a person chooses options based on which option has the largest expected utility (Dawes, 1998; Fishburn, 1981, 1999; Gauthier, 1975; Kahneman, 1994; McFadden, 1999; Resnik, 1987; Starmer, 2000; Wu, Zhang, & Gonzalez, 2004). It has been proven through several formal analyses that if people's preferences follow certain logical patterns (the so-called axioms of choice) then they are behaving as if they are maximising utility (Bermudez, 2009; Edwards, 1954; Gilboa, 2010; Jeffrey, 1983; Luce & Raiffa, 1957; Savage, 1954; von Neumann & Morgenstern, 1944). The standard view of so-called "rational man" in economics assumes such maximisation. That is, it is traditionally assumed that people have stable, underlying preferences for each of the options presented in a decision situation. It is assumed that a person's

preferences for the options available for choice are complete, well ordered, and well behaved in terms of the axioms of choice (transitivity, etc.).

The axiomatic approach to choice, because it defines instrumental rationality as adherence to certain types of consistency and coherence relationships (e.g., Luce & Raiffa, 1957; Savage, 1954), allows the rationality of nonhuman animals to be assessed as well as that of humans (Kacelnik, 2006). In fact many animals appear to have a reasonable degree of instrumental rationality, because it has been established that the behaviour of many nonhuman animals does in fact follow pretty closely the axioms of rational choice (Alcock, 2005; Arkes & Ayton, 1999; Dukas, 1998; Fantino & Stolarz-Fantino, 2005; Hurley & Nudds, 2006; Kagel, 1987; Real, 1991; Schuck-Paim & Kacelnik, 2002; Schuck-Paim, Pompilio, & Kacelnik, 2004; Stephens & Krebs, 1986).

The adaptively shaped behaviour of nonhuman animals can, in theory, deviate from the axioms of rational choice because it is possible for the optimisation of fitness at the genetic level to dissociate from optimisation at the level of the organism (Barkow, 1989; Cooper, 1989; Dawkins, 1982; Houston, 1997; Over, 2002; Skyrms, 1996; Stanovich, 2004). Although such deviations do occur in the animal world (Bateson, Healy, & Hurly, 2002, 2003; Shafir, 1994), it is nonetheless true that, as Satz and Ferejohn (1994) have noted: ''pigeons do reasonably well in conforming to the axioms of rational-choice theory'' (p. 77).

Some theorists have seen a paradox here. They point to the evidence indicating that humans often violate the axioms of rational choice (Bazerman, Tenbrunsel, & Wade-Benzoni, 1998; Kahneman, 2011; Kahneman & Tversky, 1979, 2000; Lichtenstein & Slovic, 2006; Loomes, Starmer, & Sugden, 1991; McFadden, 1999; Shafir & LeBoeuf, 2002; Shafir & Tversky, 1995; Stanovich, 2009, 2011). Humans even make errors of reasoning—honouring sunk costs for example—that animals tend to avoid (see Arkes & Ayton, 1999). Some investigators (often critics of experiments showing human error) have thought that there was something puzzling or even incorrect about these experiments indicating failures of instrumental rationality in humans, particularly in light of the findings of high levels of instrumental rationality in lower animals such as bees and pigeons. For example Gigerenzer (1994) puzzles over the fact that ''bumblebees, rats, and ants all seem to be good intuitive statisticians, highly sensitive to changes in frequency distributions in their environments... One wonders, reading that literature, why birds and bees seem to do so much better than humans'' (p. 142).

However, it is wrong to assume that rationality should increase with the complexity of the organism under study. To the contrary, there is nothing paradoxical at all about lower animals demonstrating more instrumental rationality than humans, because the principles of rational choice are actually *easier* to follow when the cognitive architecture of the organism is simpler.

One reason there is no paradox here—that it is unsurprising that bees show more instrumental rationality than humans—is that the axioms of rational choice all end up saying, in one way or another, that choices should not be affected by context. As mentioned previously, the axioms of choice operationalise the idea that an individual has pre-existing preferences for all potential options that are complete, well-ordered, and stable. When presented with options the individual simply consults the stable preference ordering and picks the one with the highest personal utility. Because the strength of each preference—the utility of that option—exists in the brain before the option is even presented, nothing about the context of the presentation should affect the preference, unless the individual judges the context to be important (to change the option in some critical way). This general property of rational choice—context independence—actually makes the strictures of rational preference more difficult to follow for more complex organisms.[1]

CONTEXT AND THE AXIOMS OF RATIONAL CHOICE

Consider a principle of rational choice called the independence of irrelevant alternatives discussed by Sen (1993). The principle of independence of irrelevant alternatives (more specifically, property alpha) can be illustrated by the following humorous imaginary situation. A diner is told by a waiter that the two dishes of the day are steak and pork chops. The diner chooses steak. Five minutes later the waiter returns and says, ''Oh, I forgot, we have lamb as well as steak and pork chops.'' The diner says, ''Oh, in that case, I'll have the pork chops.'' The diner has violated the property of independence of irrelevant alternatives, and we can see why this property is a foundational principle of rational choice by just noting how deeply odd the choices seem. Formally, the diner has chosen x when presented with x and y, but prefers y when presented with x, y, and z.

Yet consider a guest at a party faced with a bowl with one apple in it. The individual leaves the apple—thus choosing nothing (x) over an apple (y). A few minutes later, the host puts a pear (z) in the bowl. Shortly thereafter, the guest takes the apple. Seemingly, the guest has just done what the diner did in the previous example. The guest has chosen x when presented with x and y, but has chosen y when presented with x, y, and z. Has the independence of irrelevant alternatives been violated? Most will think not. Choice y in the second situation is not the same as choice y in the first—so the equivalency

[1]The same principle suggests that when pure consistency is the definition of rationality, people primed to use only Type 1 processing (Evans, 2008, 2010) might achieve better results than those using a mixture of Type 1 and Type 2 processing (see Lee, Amir, & Ariely, 2009; Reyna, 2004; Wilson & Schooler, 1991).

required for a violation of the principle seems not to hold. While choice y in the second situation is simply ''taking an apple'', choice y in the first is contextualised and probably construed as ''taking the last apple in the bowl when I am in public'' with all of its associated negative utility inspired by considerations of politeness.

What has been illustrated here is that sometimes it does make sense to contextualise a situation beyond the consumption utilities involved—but sometimes it does not.[2] In the first example it did not make sense to code the second offer of y to be ''pork chops when lamb is on the menu'' and the first to be ''pork chops without lamb on the menu''. A choice comparison between steak and pork should not depend on what else is on the menu. Sometimes though, as in the second example, the context of the situation *is* appropriately integrated with the consumption utility of the object on offer. It makes social sense, when evaluating the utilities involved in the situation, to consider the utility of the first y to be: the positive utility of consuming the apple plus the negative utility of the embarrassment of taking the last fruit in the bowl.

This example illustrates one way in which human rationality is not like animal rationality—humans code into the decision options much more contextual information. Nonhuman animals are more likely to respond on the basis of objective consumption utility without coding the nuanced social and psychological contexts that guide human behaviour. This of course is not to deny that many animals can respond to contingencies in their social environments, which is obviously true. For example, a chimp might well refrain from taking a single apple in the presence of a dominant individual (although some animals might not refrain—many of our pets get bitten by other animals for making just this mistake). The only claim being made here is that humans process more, and more complex, contextual information than other animals.

It is not difficult to demonstrate that humans incorporate a host of psychological, social, and emotional features into the options when in a choice situation—for example, see the voluminous work on the Ultimatum Game[3] (Camerer & Fehr, 2006). The problem is that the

[2]In philosophical discussions the issue of how the alternatives should be contextualised is termed the problem of the eligibility of interpretations (see Hurley, 1989, pp. 55–106) or the construal problem (Stanovich, 1999, Ch. 4).

[3]Of course the joke here is that if nonhuman animals could understand the instructions of the experiment, they would be nearly human and probably respond as humans do. Nonetheless it should be noted that, in repeated iterations of such a game, it is an open question whether any higher primate might be able to learn to use punishing refusals to shape the partner's response. Jensen, Call, and Tomasello (2007) set up a version of the Ultimate Game for chimpanzees and claimed that, unlike humans, the chimpanzees showed maximising behaviour and thus were not sensitive to fairness. Knoch, Pascual-Leone, Meyer, Treyer, and Fehr (2006) found that, in humans, disrupting the right dorsolateral prefrontal cortex (DLPFC) with repetitive transcranial magnetic stimulation resulted in more acceptances of low unfair offers. Their finding suggests that the inhibitory powers of the DLPFC are used to suppress the response that would represent

principle of the independence of irrelevant alternatives (Sen, 1993) illustrates an important fact that will be elaborated here—that the classic rational choice axioms all preclude context effects of one kind or another. Recall that that principle states that if x is chosen from the choice set x and y, then y cannot be chosen when the choice set is widened to x, y, and z. When x and y are *truly* the same across situations, one should not switch preferences from x to y when alternative z is added. Likewise, all of the other principles of rational choice have as implications, in one way or another, that irrelevant context should not affect judgement. Take a very basic principle of rational choice, transitivity (if you prefer A to B and B to C, then you should prefer A to C). The principle contains as an implicit assumption that you should not contextualise the choices such that you call the "A" in the first comparison "A in a comparison involving B" and the "A" in the third comparison "A in a comparison involving C".

Other axioms of rational choice have the same implication—that choices should be not be inappropriately contextualised (see Broome, 1990; Schick, 1987; Tan & Yates, 1995; Tversky, 1975). Consider another axiom from the theory of utility maximisation under conditions of risk, the so-called independence axiom (a different axiom from the independence of irrelevant alternatives, and sometimes termed substitutability or cancellation; see Baron, 1993; Broome, 1991; Luce & Raiffa, 1957; Neumann & Politser, 1992; Savage, 1954; Shafer, 1988; Slovic & Tversky, 1974; Tversky & Kahneman, 1986). The axiom states that if the outcome in some state of the world is the same across options, then that state of the world should be ignored. Again, the axiom dictates a particular way in which context should be disregarded. And just like the independence of irrelevant alternatives example, humans sometimes violate it because their psychological states are affected by just the contextual feature that the axiom says should not be coded into their evaluation of the options. The famous Allais (1953) paradox provides one such example. Allais proposed the following two choice problems:

Problem 1. Choose between:

A: One million dollars for sure
B: .89 probability of one million dollars
.10 probability of five million dollars
.01 probability of nothing

narrow instrumental rationality in the task. The DLPFC appears to be used to implement a goal of fairness—one that sacrifices utility maximisation narrowly construed. Lakshminarayanan and Santos (2009) claim to have shown that capuchin monkeys possess at least some of the inhibitive capacities that the Ultimatum Game requires.

Problem 2. Choose between:

C: .11 probability of one million dollars
.89 probability of nothing
D: .10 probability of five million dollars
.90 probability of nothing

Many people find option A in Problem 1 and option D in Problem 2 to be the most attractive, but these choices violate the independence axiom. To see this we need to understand that .89 of the probability is the same in both sets of choices (Savage, 1954). In both Problem 1 and Problem 2, in purely numerical terms, the participant is essentially faced with a choice between .11 probability of $1,000,000 versus .10 probability of $5,000,000 and .01 probability of nothing. If you chose the first in Problem 1 you should choose the first in Problem 2. That is, options A and C map into each other, as do B and D. The choices of A and D are incoherent.

Many theorists have analysed why individuals finding D attractive might nonetheless be drawn to option A in the first problem (Bell, 1982; Loomes & Sugden, 1982; Maher, 1993; Schick, 1987; Slovic & Tversky, 1974). Most explanations involve the assumption that the individual incorporates psychological factors such as regret into their construal of the options. But the psychological state of regret derives from the part of the option that is constant and thus, according to the axiom, should not be part of the context taken into account. For example, the zero-money outcome of option B might well be coded as something like "getting nothing when you passed up a sure chance of a million dollars!" The equivalent .01 slice of probability in option D is folded into the .90 and is not psychologically coded in the same way. Whether this contextualisation based on regret is a justified contextualisation has been the subject of intense debate (Broome, 1991; Maher, 1993; Schick, 1987; Slovic & Tversky, 1974; Tversky, 1975). Unlike the case of "taking the last apple in the bowl when I am in public", in the Allais paradox it is less clear that the .01 segment of probability in the B option should be contextualised with the negative utility of an anticipated psychological state that derives from the consequences in an outcome that did not obtain.

My point here is not to settle the debate about the Allais paradox, which has remained deadlocked for decades. Instead the point is to highlight how humans recognise subtle contextual factors in decision problems that complicate their choices (Stewart, 2009) and perhaps contribute to instability in preferences—instability that sometimes contributes to the production of a sequence of choices that violates one of the coherence constraints that define utility maximisation under the axiomatic approach of Savage (1954) and von Neumann and Morgenstern (1944). It is important to note that an agent with a less-subtle psychology might be less prone to be

drawn into complex cogitation about conflicting psychological states. An agent impervious to regret might be more likely to treat the A vs B and C vs D choices in the Allais problem as structurally analogous. Such a psychologically impoverished agent would be more likely to adhere to the independence axiom and thus be judged as instrumentally rational.

As a final example, consider the axiom of utility theory that is termed the reduction of compound lotteries—an axiom that people can also sometimes violate (Kahneman & Tversky, 1979). Imagine that we are to flip a coin and if it comes up heads you win $100 and if it comes up tails you must flip again. If, on the second flip, heads comes up you lose $25, and if tails comes up you lose $150. The reduction of compound lotteries principle states that you must consider this game to be exactly equal to a one-shot gamble in which you had 50% chance of winning $100, 25% chance of losing $25, and 25% chance of losing $150. In other words, you should consider only the final states of compound gambles and should evaluate only those final states, ignoring how the final states came about (whether by a one-shot or two-shot gamble). The axiom seems fairly straightforward, but again it involves abstracting away certain contextual features that humans might deem important. Luce and Raiffa (1957) highlight this decontextualising aspect of the compound lottery postulate when discussing their axiomatisation of utility theory: "The assumption seems quite plausible. Nonetheless, it is not empty, for it abstracts away all 'joy in gambling,' 'atmosphere of the game,' 'pleasure in suspense,' and so on, for it says that a person is indifferent between a multistage lottery and the single stage one" (p. 26).

My point in considering the principles of independence of irrelevant alternatives, transitivity, independence, and reduction of compound lotteries is to highlight one common feature of these axioms: they all require the decision maker to abstract away aspects in the contextual environment of the options. This is true as well of other principles of rational choice not discussed here such as descriptive and procedural invariance (Arrow, 1982; Kahneman & Tversky, 1984; Tversky & Kahneman, 1981, 1986). It is this fact, combined with one other assumption, that explains why it may actually be harder for humans to fulfil the strictures of instrumental rationality than lower animals. Humans are the great social contextualisers. We respond to subtle environmental cues in the contextual surround and are sensitive to social flux and nuance. All of this means that the contextual features humans code into options may lack stability both for good reasons (the social world is not stable) and bad reasons (the cues are too many and varying to be coded consistently each time).

In having more capacity for differential coding of contextual cues from occasion to occasion, humans create more opportunities for violation of any number of choice axioms, all of which require a consistent contextualisation of the options from choice to choice. The more such contextual cues are

coded, the more difficult it will be to consistently contextualise from decision to decision. The very complexity of the information that humans seek to bring to bear on a decision is precisely the thing that renders difficult an adherence to the consistency requirements of the choice axioms.[4]

SYMBOLIC COMPLEXITY AND EXPRESSIVE RATIONALITY: ''IT'S A MEANING ISSUE, NOT A MONEY ISSUE''

Another reason that human rationality is not an extension of chimpanzee rationality is because symbolic utility plays a critical role in human rational judgement, but there is no counterpart in chimpanzee rationality. The representational abilities of humans make possible a level of symbolic life unavailable to any other animal. For example, Hargreaves Heap (1992) argues for distinguishing what he terms expressive rationality from instrumental rationality. When engaged in expressively rational actions, agents are attempting to articulate and explore their values rather than trying to fulfil them. They are engaging in expressions of their beliefs in certain values, monitoring their responses to these expressions, and using this recursive process to alter and clarify their desires. Such exploratory actions are inappropriate inputs for a cost–benefit calculus that assumes fully articulated values and a single-minded focus on satisfying first-order preferences (see Anderson, 1993).

The best extant discussion of expressive rationality is contained in Nozick's (1993) treatment of what he terms symbolic utility. Nozick (1993) defines a situation involving symbolic utility as one in which an action (or one of its outcomes) ''symbolises a certain situation, and the utility of this symbolised situation is imputed back, through the symbolic connection, to the action itself'' (p. 27). Nozick notes that we are apt to view a concern for symbolic utility as irrational. This is likely to occur in two situations. The first is where the lack of a causal link between the symbolic action and the actual outcome has become manifestly obvious, yet the symbolic action continues to be performed. Nozick mentions various anti-drug measures as possibly falling in this category. In some cases, evidence has accumulated to indicate that an anti-drug programme does not have the causal effect of reducing actual drug use, but the programme is continued because it has become the *symbol* of our concern for stopping drug use. In other cases the symbolic acts will look odd or irrational if one is outside the networks of symbolic connections that give them meaning and expressiveness. Nozick

[4]Msetfi, Murphy, Simpson, and Kornbrot (2005) describe a study in which control participants displayed less-rational response patterns in a contingent judgement task than did schizophrenic participants because the latter were less prone to process contextual features of the experiment (see also Bachman & Cannon, 2005; Sellen, Oaksford, & Gray, 2005).

(1993) mentions concerns for "proving manhood" or losing face as being in this category.

Although it would be easy to classify many instances of acts carried out because of symbolic utility as irrational because of a lack of causal connection to the outcome actually bearing the utility, or because the network of social meanings that support the symbolic connection are historically contingent, Nozick warns that we need to be cautious and selective in removing symbolic actions from our lives. For example, the concern with "being a certain type of person" is a concern for living a life that embodies values that do not directly deliver utilities but are indicative of things that do. Thus I may be fully aware that performing a particular act is characteristic of a certain type of person but does not contribute causally to my becoming that type of person. But in symbolising the model of such a person, performing the act might enable me to maintain an *image* of myself. A reinforced image of myself as "that kind of person" might make it easier for me to perform the acts that *are* actually causally efficacious in making me that type of person. Thus the pursuit of symbolic utility that maintains the self-image *does* eventually get directly cashed out in terms of the results of actions that are directly causally efficacious in bringing about what I want—to be that sort of person.

For many of us the act of voting serves just this symbolic function. Many of us are aware that the direct utility we derive from the influence of our vote on the political system (a weight of one millionth or one hundred thousandth depending on the election) is less than the effort that it takes to vote (Baron, 1998; Quattrone & Tversky, 1984), yet all the same we would never miss an election! Voting has symbolic utility for us. It represents who we are. We are "the type of person" who takes voting seriously. Not only do we gain symbolic utility from voting, but it maintains a self-image that might actually help to support related actions that are more efficacious than a single vote in a national election. The self-image that is reinforced by the instrumentally futile voting behaviour might, at some later time, support my sending a sizable cheque to Oxfam, or getting involved in a local political issue, or pledging to buy from my local producers and to avoid the chain stores.

Nozick's (1993) concept of symbolic utility also has echoes in the notion of ethical preferences and commitments that has received discussion in the economic literature (Anderson, 1993; Hirschman, 1986; Hollis, 1992; Sen, 1977, 1987, 1999). The concept of ethical preferences, like symbolic utility, has the function of severing the link between observed choice and the assumption of instrumental maximisation in the economic literature. The boycott of non-union grapes in the 1970s, the boycott of South African products in the 1980s, and the interest in fair-trade products that emerged in the 1990s are examples of ethical preferences affecting people's choices and

severing the link between choice and the maximisation of personal welfare that is so critical to standard economic analyses.

In a series of papers on the meaning inherent in a decision, Medin and colleagues (Medin & Bazerman, 1999; Medin, Schwartz, Blok, & Birnbaum, 1999) have emphasised how decisions do more than convey utility to the agent but also send meaningful signals to other actors and symbolically reinforce the self-concept of the agent. Medin and Bazerman (1999, p. 541) point out that the Recipient in the Ultimatum Game who rejects a profitable offer that is considerably under 50% of the stake may be signalling that he sees positive value in punishing a greedy Allocator. Additionally he may be engaging in the symbolic act of signalling (either to himself or to others) that he is not the kind of person who condones greed. Medin and Bazerman (1999) discuss a number of experiments in which participants are shown to be reluctant to trade and/or compare items when protected values are at stake (see Anderson, 1993; Baron & Leshner, 2000; Bartels & Medin, 2007). For example, people do not expect to be offered market transactions for their pet dog, for land that has been in the family for decades, or for their wedding rings. Among Medin and Bazerman's (1999) partiicpants a typical justification for viewing such offers as insults was that "it's a meaning issue, not a money issue".

All of the symbolic complexity discussed in this section (see also Baron & Spranca, 1997, and Dehghani et al., 2010, on protected values) leads to problems in maintaining instrumental rationality in the same manner that contextual complexity does. To the extent that options are evaluated partially in terms of symbolic utility, then social context will importantly affect responses (Nozick, 1993, for example, emphasises the socially created nature of symbolic utility). Assuming that the social cues that determine symbolic utility will be complex and variable, as in the manner of the contextual complexity problem, such variability could well create inconsistency that disrupts the coherence relationships that define instrumental rationality on the axiomatic view (Luce & Raiffa, 1957; Savage, 1954).

THE STRONG EVALUATOR STRUGGLE: EVALUATING OUR PREFERENCES

Meaning-based decisions and ethical preferences do more than just complicate first-order preferences (turning "I prefer an orange to a pear" into "I prefer a Florida orange to a pear and a pear to a South African orange"). On the narrow instrumental view, the role of reason is only to serve unanalysed first-order desires. However, as Nozick (1993) says, "if human beings are simply Humean beings, that seems to diminish our stature. Man is the only animal not content to be simply an animal... It is symbolically important to us that not all of our activities are aimed at

satisfying our given desires'' (p. 138). Nozick (1993) suggests that having symbolic utilities is the way to rise above what he terms the ''Humean nexus'' (pre-existing desires in correct causal connection with actions in the instrumentally rational agent).

Notions of rationality that are caught in the Humean nexus allow for no programme of cognitive reform. Humean instrumental rationality takes a person's presently existing desires as given. In contrast, expressive actions can be used to aid in the creation of a self-image that in turn is a mechanism to be used in future goal regulation. Much in the way that cognitive scientists have speculated about auditory self-stimulation leading, in our evolutionary history, to the development of cognitive pathways between brain modules (see Dennett, 1991, 1996), the feedback from expressive actions can also help to shape the structure of goals and desires. However, such feedback can also destabilise first-order preferences in ways that make choices more likely to violate the consistency criteria of rational choice.

Philosophers have long stressed the importance of self-evaluation in the classically modernist quest to find one's truest personal identity. Taylor (1989), for example, stresses the importance of what he terms strong evaluation, which involves ''discriminations of right or wrong, better or worse, higher or lower, which are not rendered valid by our own desires, inclinations, or choices, but rather stand independent of these and offer standards by which they can be judged'' (p. 4). There is a philosophical literature on the notion of strong evaluation (Bratman, 2003; Dworkin, 1988; Harman, 1993; Lehrer, 1990, 1997; Lewis, 1989; Maher, 1993; Watson, 1975), and it is one that is of potential theoretical interest for decision scientists (see Flanagan, 1996, for an insightful discussion of Taylor, 1989, that is informed by cognitive science). Taylor's (1989) notion of strong evaluation can be a bit more formally explicated in language more commonly used by economists, decision theorists, and cognitive psychologists (see Jeffrey, 1974; Kahneman & Tversky, 2000; Slovic, 1995; Tversky, Slovic, & Kahneman, 1990). What Taylor calls strong evaluation would be termed, in these other literatures, a second-order preference: a preference for a particular set of first-order preferences.

In a classic paper on second-order desires, Frankfurt (1971) speculated that only humans have such meta-representational states. He evocatively termed creatures without second-order desires (other animals, human babies) ''wantons''. To say that a wanton does not form second-order desires does not mean that they are heedless or careless about their first-order desires. Wantons can be rational in the thin, purely instrumental, sense. Wantons may well act in their environments to fulfil their goals with optimal efficiency. A wanton simply does not reflect upon his/her goals. A strong evaluator, in contrast, engages in a higher-order evaluation of the first-order preferences (see Velleman, 1992).

So, for example, using the preference relationship that is the basis for the formal axiomatisation of utility theory, we can illustrate the situation. If John prefers to smoke, we have:

S pref $\sim$S

However, humans alone appear to be able to represent a model of an idealised preference structure—perhaps, for example, a model based on a superordinate judgement of long-term lifespan considerations (or what Gauthier, 1986, calls considered preferences). So a human can say: I would prefer to prefer not to smoke. Only humans can decouple from a first-order desire and represent:

($\sim$S pref S) pref (S pref $\sim$S)

This second-order preference then becomes a motivational competitor to the first-order preference. At the level of second-order preferences, John prefers to prefer to not smoke; nevertheless, as a first-order preference, he prefers to smoke. The resulting conflict signals that John lacks what Nozick (1993) terms rational integration in his preference structure.[5]

Importantly, such a mismatched first-order/second-order preference structure is one reason why humans are often less rational than bees in an axiomatic sense (see Stanovich, 2004). This is because the struggle to achieve rational integration can destabilise first-order preferences in ways that make them more prone to violate the consistency requirements of the basic axioms of utility theory.

Engaging in a second- (or higher-) order critique of first-order preferences might actually mean temporarily sacrificing some degree of instrumental rationality because of the desire to seek rational integration across all vertical levels of preference. Any instrumental loss caused by

[5]There of course is no limit to the hierarchy of higher-order desires that might be constructed. But the representational abilities of humans may set some limits— certainly three levels seems a realistic limit for most people in the nonsocial domain (Dworkin, 1988). However, third-order judgements can be called upon to help achieve rational integration at lower levels. So, for example, John, the smoker, might realise when he probes his feelings that: He prefers his preference to prefer not to smoke over his preference for smoking:

[($\sim$S pref S) pref (S pref $\sim$S)] pref [S pref $\sim$S]

We might in this case say that John's third-order judgement has ratified his second-order strong evaluation. Presumably this ratification of his second-order judgement adds to the cognitive pressure to change the first-order preference by taking behavioural measures that will make change more likely (entering a smoking secession programme, consulting his physician, asking the help of friends and relatives, staying out of smoky bars, etc.).

instability in the first-order preferences thus induced is the result of an attempt to engage in the broader cognitive programme of critiquing lower-level desires by forming higher-level preferences. The instrumental rationality of nonhuman animals is not threatened by any such destabilising programme.

ESCAPING THE RATIONALITY OF CONSTRAINT: CONTEXTUAL COMPLEXITY, SYMBOLIC COMPLEXITY, AND THE STRONG EVALUATOR STRUGGLE

Thus, as with the other two mechanisms discussed—contextual complexity and symbolic complexity—the strong evaluator struggle leads human behaviour to deviate from instrumental rationality. Many authors have commented on how the behaviour of entities in very constrained situations (firms in competitive markets, people living in subsistence-agriculture situations, animals in predator-filled environments) are the entities whose behaviours fit the rational choice model the best (e.g., Clark, 1997, pp. 180–184; Denzau & North, 1994; Satz & Ferejohn, 1994). The harshly simplified environments of these entities allow only for the valuing of instrumental rationality in the most narrowly defined way (or else the entities perish). These environments are all subject to evolutionary or quasi-evolutionary (e.g., markets) selection processes. Only entities that fit the narrow criteria of rational choice are around to serve as subjects of study. Tiny charitable contributions aside, corporations are not notable for valuing symbolically (any corporation valuing symbolic acts more than the bottom line would be rapidly eliminated in the market); nor do they engage in a struggle between their desire for profits and some higher-order preference. Corporations, like animals in harsh environments, achieve what we might call the instrumental rationality of constraint. Economic theorists who stress the rationality of humans tend to analyse situations where there is no choice; or, more specifically, they analyse situations set up to ruthlessly exploit those not making the instrumentally optimal choice.

Most humans now do not operate in such harsh selective environments of constraint (outside many work environments that deliberately create constraints, such as markets). They use that freedom to pursue symbolic utility, thereby creating complex, context-dependent preferences that are more likely to violate the strictures of coherence that define instrumental rationality. But those violations do not make them inferior to the instrumentally rational pigeon. Degrees of rationality among entities pursuing goals of differing complexity are not comparable. One simply cannot count up the number of violations and declare the entity with fewer violations the more rational. The degree of instrumental rationality achieved must be contextualised according to the complexity of the goals pursued.

In addition to pursuing symbolic rationality, humans engage in the risky project of evaluating their desires by forming higher-order preferences and examining whether they rationally cohere with their first-order preferences. This is a risky project because the potential lack of rational integration (conflicts between first- and higher-order preferences) thereby entailed puts instrumental rationality in jeopardy. Optimisation of first-order desire satisfaction might not proceed optimally as long as this programme of cognitive criticism and rational integration continues. This cost in instrumental rationality is the price humans pay for being a species, the only species, that cares about what it cares about (Frankfurt, 1982; Penn, Holyoak, & Povinelli, 2008).

We are the only species that disrupts the coherence of its preferences by destabilising them through cognition directed at self-improvement and self-determination. This is because we are the only species that engages in true Type 2 processing (Evans & Stanovich, in press). I am of course referring here to dual-process models of mind, now enjoying a resurgence in psychology (Evans, 2003, 2008; Kahneman, 2011; Stanovich, 2004, 2011). Such models capture a phenomenal aspect of human decision making that is little commented upon, yet is of profound importance—that humans often feel alienated from their choices. This feeling of alienation, although emotionally discomfiting when it occurs, is actually a reflection of a unique aspect of human cognition—the use of the meta-representational abilities of the analytic system to enable a cognitive critique of our beliefs and our desires (Nichols & Stich, 2003; Perner, 1991; Sperber, 2000; Stanovich, 2004).

These meta-representational abilities include the ability to decouple our mental representations from the world so that they can be reflected upon and potentially improved. A unique aspect of Type 2 thinking in dual-process theories (Evans & Stanovich, in press), these decoupling operations allow us to mark a belief as a hypothetical state of the world rather than a real one (e.g., Carruthers, 2002; Dienes & Perner, 1999; Evans & Over, 2004). Decoupling abilities prevent our representations of the real world from becoming confused with representations of imaginary situations that we create on a temporary basis. Thus decoupling processes enable one to distance oneself from representations of the world (or from goal states) so that they can be reflected upon and potentially improved.

META-RATIONALITY: QUESTIONING THE APPLICABILITY OF RATIONAL PRINCIPLES

Humans are also unique in recognising the importance of normative rules while at the same time realising that such rules are open to critique. Consider the example of sunk costs (Thaler, 1980). The traditional rational stricture is that they should be ignored. Decisions should concern only

future consequences, and sunk costs are in the past. So if one would turn off the movie if it were free, then one should also turn off the movie and do something else if one had already paid $7 for it. You should not continue to do something that in the future would make you less happy because you have spent the money in the past. But Keys and Schwartz (2007) point out that there seems nothing wrong with feeling that the memory that you had paid the $7 might depress the enjoyment of whatever else you decided to do. You might feel bad because you had "thrown money down the drain" by not watching. Whatever the normative status of the principle of ignoring sunk costs, it seems right for people to think that "not watching the movie and regretting that you spent the $7" is a worse outcome than that of "not watching the movie". Why shouldn't people take into account the regret that they will feel if they fail to honour sunk costs. Keys and Schwartz (2007) introduce the concept of "leakage" to help us understand this situation. Traditionally we differentiate the decision itself from the experience of the consequences of that decision. At the time of the decision, the $7 already spent should not be a factor—so says the principle of ignoring sunk costs. But what if that $7 already spent will in fact affect your experience of one of the alternatives (here, specifically, the experience of the alternative of turning off the movie). If so, then the effect of the $7 (the regret at having "wasted" it) has functionally leaked into the experience of the consequences—and if it is in fact part of the consequence of that option, then why should it be ignored?

Keys and Schwartz (2007) cite studies where (seemingly) irrelevant factors that are used to frame the choice actually leak into the experience of the alternative chosen. They cite studies in which beef labelled as "75% lean" was experienced as tasting better than beef labelled "25% fat" and in which people performed better after drinking an energy drink that cost $1.89 than after consuming the same drink when it cost only $0.89. Again, the way the alternatives were framed leaked into the *experience* of the alternatives. So one argument is that framing effects in such situations are not irrational because the frame leaks into the experience of the consequence. In fact, Keys and Schwartz (2007) point out that if leakage is a fact of life, then the wise decision maker might actually want to take it into account when making decisions.

Consider an example of regret as a leakage factor. Keys and Schwartz (2007) discuss the example of standing in the grocery store line and suspecting that the neighbouring line would move faster. What should we do, stay or switch? On the one hand, our visual inspection of the neighbouring line and the people in it leads us to suspect that it would move faster. Why would we ever hesitate if this were our judgement? Often the reason we do in fact hesitate is because we can recall instances in the past where we have switched lines and then observed that our original line actually ended up moving faster. We want to kick ourselves when this happens—we regret our decision to

switch. And we tend to regret it more than when we fail to switch and the neighbouring line does indeed move faster. If we take this anticipatory regret into account we might well decide to stay in the line we are in even when the neighbouring line looks like it will move faster.

In the grocery line and the movie examples anticipated regret leads us to take actions that would otherwise not be best for us (in the absence of such anticipation). One response to these choices is to defend them as rational cases of taking into account aspects of decision framing that actually do leak into the experienced utility of the action once taken. Another response is to argue that if regret is leading us away from actions that would otherwise be better for us, then perhaps what should be questioned is whether the regret we feel in various situations is appropriate.

This response—that maybe we should *not* let aspects of how the choices are framed leak into our experience—Keys and Schwartz call "leak plugging". That leak plugging may sometimes be called for is suggested by another example that they discuss—that students think that if they change their responses on a multiple-choice test that they are more likely to change a correct response into an incorrect one than vice versa. Keys and Schwartz point out that this belief is false, but that it may be a superstition that arose to help prevent regret. That there is another response to regret other than avoidance is suggested by a question that we might ask about the situation surrounding the multiple-choice superstition: Are people better off with lower grades and reduced regret, or are they better off with some regret but higher grades?

The multiple choice example thus suggests another response to decision leakage of contextual factors—that rather than simply accommodating such leakage into our utility calculations, we might consider getting rid of the leakage. In short, maybe the most rational thing to do is to condition ourselves to avoid regret in situations where we would choose otherwise without it. Without the regret we could freely and rationally choose to turn off the movie and enjoy an activity that is more fulfilling than watching a boring film. Without the regret we could change to whichever grocery line looked more promising and not worry about our feelings if our predicted outcome did not occur.

Note that a decision to condition ourselves to avoid regret in the movie example would represent a more critical use of the rational principle of avoiding the honouring of sunk costs. It would reflect a use of the sunk cost principle that was informed by a meta-rational critique—one that took a critical stance towards the rational principle rather than applying it blindly. A first-order use of the sunk cost principle would apply it no matter what and—given the natural structure of human psychology—would sometimes result in lower experienced utility because the blind use of the principle fails to account for regret. A critical stance towards the principle would recognise

that sometimes it leads to lower experienced utility due to the unaccounted-for regret. But, as a further step in meta-rational analysis, the regret itself might be critiqued. The sunk cost principle comes into play again in reminding us that, absent the regret, turning off the movie is the better choice. If, at this point, we decide to endorse the sunk cost principle, it is in a much more reflective way than simply blindly applying it as a rule without a consideration of human psychology. The decision to alter our psychologies in light of the rule would in a sense be a second-order use of the rule, one that represented a meta-rational judgement.

This aspect of meta-rationality is in effect asking about the appropriateness of our emotional reactions to a decision. If we deem these reactions appropriate then they must be factored in. Sometimes, however, we will deem the emotions less important than our other goals. We will want the better grades, the better line at the grocery store, and the activity that is better than the boring movie—and we will want all of these things more than we value avoiding regret. In this case we revert to the traditional normative rule—but only after having engaged in meta-rational reflection.

Keys and Schwartz (2007) discuss how situations that are repeated are more likely to be the ones where we might want to plug leakage and target some of our emotions for reform. Someone afraid of elevators might be perfectly rational, on a particular occasion, in taking the stairs even though the stairs are slower, because they have factored in the negative utility of their fear while riding in the elevator. However, such a person living and working in New York City might well think of accepting some therapy in the service of ridding themselves of this fear. What might look rational on a given single occasion might seem very suboptimal from the standpoint of a *lifespan* filled with similar activities. Financial decisions that cumulate have a similar logic. Suppose you are the type of person who is affected by friendly salespeople. You tend to buy products from those who are friendly. Furthermore, suppose that there is leakage from decision to experience regarding this factor—you actually enjoy products more when you have purchased them from friendly people. Clearly though, given the logic of our market-based society, you are going to end up paying much more for many of your consumer goods throughout your lifetime. Here a lifetime and a single case tend to look very different. You pay 25 cents more for a coffee from the Bean People tomorrow because you like them better than the Java People. No problem. But you might answer differently if calculations were to show that buying from friendly people will cost you a compounded return of $175,667 in your retirement fund over a lifetime. With this information, you might decide to plug the leakage and stop responding to the "friendly factor" in your future decisions.

An actual consumer example comes from the "extended warranties" that are sold with many appliances. At the time of each individual purchase these

small-scale insurance contracts may give us some reassurance and comfort. But consumer magazines routinely report that, when aggregated, these are very bad products. That is, across a number of such contracts, the return to the consumer is very low—much more is spent in premiums than is actually returned by making a claim on the warranty. Of course, on one particular purchase, buying the warranty might have positive utility—not because of any money saved but because it reduces the negative utility of the anxiety we feel at the time of purchase. Nonetheless, however comforting the warranty is in the case *this particular* appliance, across several such appliances they are a very bad deal. Thus the consumer is better off by trying to get rid of the purchase anxiety that leads them to buy the warranty each time.

These examples show the more delicate interplay between normative rules, individual decisions, and a long-term view of one's goals and desires that takes place when meta-rationality rather than a thin instrumental rationality is our concern. It is a mistake to apply a blanket normative rule too quickly to a specific situation that may have alternative interpretations and subtle contextual factors that might leak into the experience of the consequences. However, it is equally an error to fail to take a broader view of life—one that would examine how certain responses may have cumulative effects over time. Additionally, people often fail to recognise that a market-based society is often a hostile environment for those who have "alternative explanations" of situations—because other individuals in such societies will take steps to exploit the "alternative interpretation" of the decision maker (those who view an insurance salesperson as a financial planner are in trouble). A broader view of life, one that recognised hostile environments and that recognised the cumulative effect of repeated instances, might dictate a critique of an alternative construal even though on a one-shot basis it might seem rational.

TWO-TIERED RATIONALITY EVALUATION: A LEGACY OF HUMAN COGNITIVE ARCHITECTURE

People aspire to rationality broadly conceived (see Elster, 1983), not just a thin instrumental rationality. People want their desires satisfied, true, but they are also concerned about having the *right* desires. Because humans aspire to rationality broadly rather than narrowly defined, a two-tiered evaluation of their rationality is necessary. The instrumental rationality a person achieves must be evaluated by taking into account the complexity of the goals being pursued and by analysing the dynamics of the cognitive critique the person engages in. Or, to put it another way, both thin and broad rationality (to use Elster's terms) need evaluation.

The rules for examining instrumental rationality are well articulated. In contrast, the criteria that should be applied when evaluating broad

rationality are much more complex and contentious (see Nozick's 1993 discussion of 23 criteria for the evaluation of preferences) but would certainly include: the degree of strong evaluation undertaken; the degree to which a person finds lack of rational integration aversive and is willing to take steps to rectify it (Nozick's 1993 principle IV); whether the individual can state a reason for all second-order desires (Nozick's 1993 principle VII); whether it is the case that a person's desires are not such that acting on them leads to irrational beliefs (Nozick's 1993 principle XIV); whether a person avoids forming desires that are impossible to fulfil (Nozick's 1993 principle X), and others (see Nozick, 1993).

We can thank a feature of our cognitive architecture for making the pursuit of broad rationality possible. In holding, contemplating, and evaluating second-order desires that conflict with first-order desires, we are cognitively dealing with a hypothesised mental state—one that is actually not true of us. We are able to represent a state of affairs that does not map into an actual, causally active, mental state of our own. We are able to mark a mental state as not factual. Many cognitive theorists (Carruthers, 2006; Dienes & Perner, 1999; Evans & Over, 1999, 2004; Nichols & Stich, 2003; Sperber, 2000; Sterelny, 2001; Suddendorf & Corballis, 2007) have emphasised the critical importance (and specialness to human mentality) of being able to separate a belief or desire from its coupling to the world (to mark it as a hypothetical state). These meta-representational abilities make possible the higher-order evaluations that determine whether we are pursuing the right aims. They allow the query "if I had a different set of desires, it would be preferable to the ones I have now" and they appear to be uniquely human (Penn et al., 2008). They make it possible to add symbolic utility to our lives and actions. They are unique cognitive states, but they may well be disruptive of the coherence among preferences that defines instrumental rationality on the axiomatic view. If rationality is defined only in terms of these criteria, it is not paradoxical at all that humans will appear less rational than some nonhuman animals. No other animal is engaged in such extensive contextualisation of options, displays a concern for symbolic/expressive utility, or engages in the strong evaluator struggle.

TWO-TIERED RATIONALITY ASSESSMENT IS NOT A PANGLOSSIAN ESCAPE HATCH

It is important to realise that two-tiered rationality evaluation should not be viewed as an escape hatch for the Panglossian theorist (see Stanovich, 1999, 2004) who wishes to deny to the existence of human irrationality. The Panglossian position (of economists for example) is *not* that seeming violations of rational strictures occur because individuals are sacrificing instrumental rationality in order to engage in a programme of cognitive

critique based on higher-order preferences. Instead they posit that perfect instrumental rationality is attained *whatever* the internal and external perturbations impinge on the agent.

In the course of two-tier rationality evaluation we will no doubt find that some irrationality at the instrumental level arises because an individual is aspiring to something more than satisfying first-order desires (i.e., that the very critique of those desires sometimes disrupts the achievement of instrumental rationality by axiomatic criteria). However, not all thin-theory irrationality will derive from such factors. How can we tell how much deviation from instrumental rationality has the strong evaluator struggle as its cause? When the violations concern differential contextualisation in the way precluded by the axioms of choice, we can garner critical information from the decision maker and apply Wilson and Brekke's (1994) concept of mental contamination. Mental contamination occurs when an individual's behaviour is affected by factors that they wish were not implicated in their decisions.

Consider framing effects and preference reversals, two of the most researched ways of demonstrating deviations from thin-theory, instrumental rationality (Kahneman, 2011; Lichtenstein & Slovic, 2006). In such problems participants often agree in post-experimental interviews that the two versions are identical and that they should not be affected by the wording. This finding is somewhat embarrassing for Panglossian views that stress the multiplicity of norms. In fact people most often retrospectively endorse the Bayesian and SEU norms that they routinely violate (Shafir, 1993; Shafir & Tversky, 1995). In introducing the collection of Amos Tversky's writings, Shafir (2003) stresses this very point: "The research showed that people's judgements often violate basic normative principles. At the same time, it showed that they exhibit sensitivity to these principles' normative appeal" (p. x). For example, Koehler and James (2009) found that non-normative "probability matchers rate an alternative strategy (maximising) as superior when it is described to them" (p. 123).

In short, preference reversals or framing effects do not represent alternative contextualisations that participants *want* to have. Instead such violations of descriptive invariance represent true failures of instrumental rationality. They do not result from alternative contextualisations caused by differential symbolic utility or by instability deriving from strong evaluation at higher levels of preference. Two-tier rationality evaluation complicates assessment at the instrumental level, but it is still possible to identify thin-theory violations.

Manuscript received 22 March 2012
Revised manuscript received 2 July 2012
First published online 25 September 2012

REFERENCES

Alcock, J. (2005). *Animal behavior: An evolutionary approach (*8th ed.). Sunderland, MA: Sinauer Associates.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'école americaine. *Econometrica*, *21*, 503–546.

Anderson, E. (1993). *Value in ethics and economics*. Cambridge, MA: Harvard University Press.

Arkes, H. R., & Ayton, P. (1999). The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin*, *125*, 591–600.

Arrow, K. J. (1982). Risk perception in psychology and economics. *Economic Inquiry*, *20*, 1–9.

Bachman, P., & Cannon, T. D. (2005). Cognitive and neuroscience aspects of thought disorder. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 493–526). New York: Cambridge University Press.

Barkow, J. H. (1989). *Darwin, sex, and status: Biological approaches to mind and culture*. Toronto: University of Toronto Press.

Baron, J. (1993). *Morality and rational choice*. Dordrecht: Kluwer.

Baron, J. (1998). *Judgement misguided: Intuition and error in public decision making*. New York: Oxford University Press.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values? *Journal of Experimental Psychology: Applied*, *6*, 183–194.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*, 1–16.

Bartels, D. M., & Medin, D. L. (2007). Are morally-motivated decision makers insensitive to the consequences of their choices? *Psychological Science*, *18*, 24–28.

Bateson, M., Healy, S. D., & Hurly, A. (2002). Irrational choices in hummingbird foraging behaviour. *Animal Behaviour*, *63*, 587–596.

Bateson, M., Healy, S. D., & Hurly, A. (2003). Context-dependent foraging decisions in rufous hummingbirds. *Proceedings of the Royal Society (London) B*, *270*, 1271–1276.

Bazerman, M., Tenbrunsel, A., & Wade-Benzoni, K. (1998). Negotiating with yourself and losing: Understanding and managing conflicting internal preferences. *Academy of Management Review*, *23*, 225–241.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*, 961–981.

Bermudez, J. L. (2009). *Decision theory and rationality*. New York: Oxford University Press.

Bratman, M. E. (2003). Autonomy and hierarchy. *Social Philosophy & Policy*, *20*, 156–176.

Broome, J. (1990). Should a rational agent maximise expected utility? In K. S. Cook & M. Levi (Eds.), *The limits of rationality* (pp. 132–145). Chicago: University of Chicago Press.

Broome, J. (1991). *Weighing goods: Equality, uncertainty, and time*. Oxford, UK: Blackwell.

Camerer, C. F., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science*, *311*, 47–52.

Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, *25*, 657–726.

Carruthers, P. (2006). *The architecture of the mind*. New York: Oxford University Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Cooper, W. S. (1989). How evolutionary biology challenges the classical theory of rational choice. *Biology and Philosophy*, *4*, 457–481.

Dawes, R. M. (1998). Behavioral decision making and judgement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1) (pp. 497–548). Boston: McGraw-Hill.

Dawkins, R. (1982). *The extended phenotype*. New York: Oxford University Press.

Dehghani, M., Atran, S., Iliev, R., Sachdeva, S., Medin, D., & Ginges, J. (2010). Sacred values and conflict over Iran's nuclear program. *Judgement and Decision Making*, *5*, 540–546.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown.
Dennett, D. C. (1996). *Kinds of minds: Toward an understanding of consciousness*. New York: Basic Books.
Denzau, A. T., & North, D. C. (1994). Shared mental models: Ideologies and institutions. *Kyklos*, *47*, 3–31.
Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–808.
Dukas, R. (Ed.). (1998). *Cognitive ecology: The evolutionary ecology of information processing and decision making*. Chicago: University of Chicago Press.
Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge, UK: Cambridge University Press.
Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380–417.
Elster, J. (1983). *Sour grapes: Studies in the subversion of rationality*. Cambridge, UK: Cambridge University Press.
Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.
Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgement, and social cognition. *Annual Review of Psychology*, *59*, 255-278.
Evans, J. St. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford, UK: Oxford University Press.
Evans, J. St. B. T., & Over, D. E. (1999). Explicit representations in hypothetical thinking. *Behavioral and Brain Sciences*, *22*, 763–764.
Evans, J. St. B. T., & Over, D. E. (2004). *If*. Oxford, UK: Oxford University Press.
Evans, J. St. B. T., & Stanovich, K. E. (in press). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science*.
Fantino, E., & Stolarz-Fantino, S. (2005). Decision-making: Context matters. *Behavioural Processes*, *69*, 165–171.
Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision*, *13*, 139–199.
Fishburn, P. C. (1999). The making of decision theory. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards*. Boston: Kluwer Academic Publishers.
Flanagan, O. (1996). *Self expressions: Mind, morals, and the meaning of life*. New York: Oxford University Press.
Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5–20.
Frankfurt, H. (1982). The importance of what we care about. *Synthese*, *53*, 257–272.
Gauthier, D. (1975). Reason and maximisation. *Canadian Journal of Philosophy*, *4*, 411–433.
Gauthier, D. (1986). *Morals by agreement*. Oxford, UK: Oxford University Press.
Gigerenzer, G. (1994). Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 129–161). Chichester, UK: John Wiley & Sons.
Gilboa, I. (2010). *Rational choice*. Cambridge, MA: The MIT Press.
Hargreaves Heap, S. P. (1992). *Rationality and economics*. Cambridge, UK: Cambridge University Press.
Harman, G. (1993). Desired desires. In R. G. Frey & C. W. Morris (Eds.), *Value, welfare, and morality* (pp. 138–157). Cambridge, UK: Cambridge University Press.
Hirschman, A. O. (1986). *Rival views of market society and other recent essays*. New York: Viking.
Hollis, M. (1992). Ethical preferences. In S. Hargreaves Heap, M. Hollis, B. Lyons, R. Sugden, & A. Weale (Eds.), *The theory of choice: A critical guide* (pp. 308–310). Oxford, UK: Blackwell.

Houston, A. I. (1997). Natural selection and context-dependent values. *Proceedings of the Royal Society (London) B*, *264*, 1539–1541.
Hurley, S. (1989). *Natural reasons: Personality and polity*. New York: Oxford University Press.
Hurley, S., & Nudds, M. (Eds.). (2006). *Rational animals?* Oxford, UK: Oxford University Press.
Jeffrey, R. (1974). Preferences among preferences. *Journal of Philosophy*, *71*, 377–391.
Jeffrey, R. C. (1983). *The logic of decision (2nd ed.)*. Chicago: University of Chicago Press.
Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are rational maximisers in an ultimatum game. *Science*, *318*, 107–109.
Kacelnik, A. (2006). Meanings of rationality. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 87–106). Oxford, UK: Oxford University Press.
Kagel, C. J. (1987). Economics according to the rats (and pigeons too): What have we learned and what we hope to learn. In A. Roth (Ed.), *Laboratory experimentation in economics: Six points of view* (pp. 587–703). New York: Cambridge University Press.
Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus & Giroux.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.
Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.
Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, UK: Cambridge University Press.
Keys, D. J., & Schwartz, B. (2007). "Leaky" rationality: How research on behavioral decision making challenges normative standards of rationality. *Perspectives on Psychological Science*, *2*, 162–180.
Knoch, D., Pascual-Leone, A., Meyer, K., Treyer, V., & Fehr, E. (2006). Diminishing reciprocal fairness by disrupting the right prefrontal cortex. *Science*, *314*, 829–832.
Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*, 123–127.
Kohler, W. (1927). *The mentality of apes (2nd ed.)*. London: Routledge & Kegan Paul.
Lakshminarayanan, V., & Santos, L. R. (2009). Cognitive preconditions for responses to fairness: An object retrieval test of inhibitory control in capuchin monkeys (*Cebus apella*). *Journal of Neuroscience, Psychology, and Economics*, *2*, 12–20.
Lee, L., Amir, O., & Ariely, D. (2009). In search of homo economicus: Cognitive noise and the role of emotion in preference consistency. *Journal of Consumer Research*, *36*, 173–187.
Lehrer, K. (1990). *Theory of knowledge*. London: Routledge.
Lehrer, K. (1997). *Self-trust: A study of reason, knowledge, and autonomy*. Oxford, UK: Oxford University Press.
Lewis, D. (1989). Dispositional theories of value. Proceedings of the Aristotelian Society, Supplementary Volume *63*, 113–137.
Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge, MA: Cambridge University Press.
Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica*, *59*, 425–439.
Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*, 805–824.
Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.
Maher, P. (1993). *Betting on theories*. Cambridge, UK: Cambridge University Press.
McFadden, D. (1999). Rationality for economists? *Journal of Risk and Uncertainty*, *19*, 73–105.

Medin, D. L., & Bazerman, M. H. (1999). Broadening behavioral decision research: Multiple levels of cognitive processing. *Psychonomic Bulletin & Review*, *6*, 533–546.
Medin, D. L., Schwartz, H. C., Blok, S. V., & Birnbaum, L. A. (1999). The semantic side of decision making. *Psychonomic Bulletin & Review*, *6*, 562–569.
Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. (2005). Depressive realism and outcome density bias in contingency judgements: The effect of the context and intertrial interval. *Journal of Experimental Psychology: General*, *134*, 10–22.
Neumann, P. J., & Politser, P. E. (1992). Risk and optimality. In J. F. Yates (Ed.), *Risk-taking behavior* (pp. 27–47). Chichester, UK: John Wiley.
Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds*. Oxford, UK: Oxford University Press.
Nozick, R. (1993). *The nature of rationality*. Princeton, NJ: Princeton University Press.
Over, D. E. (2002). The rationality of evolutionary psychology. In J. L. Bermudez & A. Millar (Eds.), *Reason and nature: Essays in the theory of rationality* (pp. 187–207). Oxford, UK: Oxford University Press.
Quattrone, G., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46*, 237–248.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*, 109–178.
Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*, 980–986.
Resnik, M. D. (1987). *Choices: An introduction to decision theory*. Minneapolis: University of Minnesota Press.
Reyna, V. F. (2004). How people make decisions that involve risk. *Current Directions in Psychological Science*, *13*, 60–66.
Satz, D., & Ferejohn, J. (1994). Rational choice and social theory. *Journal of Philosophy*, *9102*, 71–87.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schick, F. (1987). Rationality: A third dimension. *Economics and Philosophy*, *3*, 49–66.
Schuck-Paim, C., & Kacelnik, A. (2002). Rationality in risk-sensitive foraging choices by starlings. *Animal Behaviour*, *64*, 869–879.
Schuck-Paim, C., Pompilio, L., & Kacelnik, A. (2004). State-dependent decisions cause apparent violations of rationality in animal choice. *PLOS Biology*, *2*, 2305–2315.
Searle, J. R. (2001). *The rationality of action*. Cambridge, MA: MIT Press.
Sellen, J., Oaksford, M., & Gray, N. S. (2005). Schizotypy and conditional reasoning. *Schizophrenia Bulletin*, *31*, 105–116.
Sen, A. (1987). *On ethics & economics*. Oxford, UK: Blackwell.
Sen, A. (1993). Internal consistency of choice. *Econometrica*, *61*, 495–521.
Sen, A. (1999). *Development as freedom*. New York: Knopf.
Sen, A. K. (1977). Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs*, *6*, 317–344.
Shafir, E. (1993). Intuitions about rationality and cognition. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and philosophical perspectives* (pp. 260–283). London: Routledge.
Shafir, S. (1994). Intransitivity of preferences in honey bees: Support for comparative-evaluation of foraging options. *Animal Behavior*, *48*, 55–67.
Shafir, E. (Ed.). (2003). *Preference, belief, and similarity: Selected writings of Amos Tversky*. Cambridge, MA: MIT Press.
Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology*, *53*, 491–517.

Shafir, E., & Tversky, A. (1995). Decision making. In E. E. Smith & D. N. Osherson (Eds.), *Thinking* (Vol. 3) (pp. 77–100). Cambridge, MA: The MIT Press.

Skyrms, B. (1996). *The evolution of the social contract*. Cambridge, UK: Cambridge University Press.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*, 364–371.

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science*, *19*, 368–373.

Sperber, D. (2000). Meta-representations in evolutionary perspective. In D. Sperber (Ed.), *Meta-representations: A multidisciplinary perspective* (pp. 117–137). Oxford, UK: Oxford University Press.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*, 332–382.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton: Princeton University Press.

Sterelny, K. (2001). *The evolution of agency and other essays*. Cambridge, UK: Cambridge University Press.

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology*, *62*, 1041–1062.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel and is it unique to humans? *Behavioral and Brain Sciences*, *30*, 299–351.

Tan, H., & Yates, J. F. (1995). Sunk cost effects: The influences of instruction and future return estimates. *Organizational Behavior and Human Decision Processes*, *63*, 311–319.

Taylor, C. (1989). *Sources of the self: The making of modern identity*. Cambridge, MA: Harvard University Press.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior and Organization*, *1*, 39–60.

Tversky, A. (1975). A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis*, *9*, 163–173.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, *59*, 251–278.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *American Economic Review*, *80*, 204–217.

Velleman, J. D. (1992). What happens when somebody acts? *Mind*, *101*, 461–481.

von Neumann, J., & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton: Princeton University Press.

Watson, G. (1975). Free agency. *Journal of Philosophy*, *72*, 205–220.

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgements and evaluations. *Psychological Bulletin*, *116*, 117–142.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*, 181–192.

Wu, G., Zhang, J., & Gonzalez, R. (2004). Decision under risk. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgement and decision making* (pp. 399–423). Malden, MA: Blackwell Publishing.