THINKING & REASONING, 2008, 14 (1), 111–127

Psychology Press
Taylor & Francis Group

Theoretical note

Higher-order preferences and the Master Rationality Motive

Keith E. Stanovich
University of Toronto, Canada

The cognitive critique of the goals and desires that are input into the implicit calculations that result in instrumental rationality is one aspect of what has been termed broad rationality (Elster, 1983). This cognitive critique involves, among other things, the search for rational integration (Nozick, 1993)—that is, consistency between first-order and second-order preferences. Forming a second-order preference involves metarepresentational abilities made possible by mental decoupling operations. However, these decoupling abilities are separable from the motive that initiates the cognitive critique itself. I argue that Velleman (1992) has identified that motive ("the desire to act in accordance with reasons"), and that it might be operationalisable as a thinking disposition at a very superordinate cognitive level. This thinking disposition, the Master Rationality Motive, is likely to be of particular importance in explaining individual differences in the tendency to seek rational integration. Preliminary research on related constructs suggests that this construct is measurable.

Multiple-process models of mind, including dual-process theories (Evans, 2003, 2006, 2007; Evans & Over, 1996, 2004; Kahneman & Frederick, 2002, 2005; Sloman, 1996; Stanovich, 1999, 2004), are now enjoying a resurgence in psychology. 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. We display what both folk psychology and philosophers term weakness of will. For example, we continue to smoke when we know that it is a harmful habit. We order a sweet

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: kstanovich@oise.utoronto.ca

Preparation of this paper was supported by grants from Social Sciences and Humanities Research Council of Canada and the Canada Research Chairs program to Keith E. Stanovich. David Over and two anonymous reviewers are thanked for the comments on an earlier version of the manuscript.

http://www.psypress.com/tar DOI: 10.1080/13546780701384621

after a large meal, merely an hour after pledging to ourselves that we will not. But we display alienation from our responses even in situations that do not involve weakness of will—we find ourselves recoiling from the sight of a disfigured person even after a lifetime of dedication to diversity and inclusion.

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 metarepresentational abilities of the analytic system to enable a cognitive critique of our beliefs and our desires. Beliefs about how well we are forming beliefs become possible because of such metarepresentation, as does the ability to evaluate one's own desires—to desire to desire differently. In this theoretical note, I focus on the latter—so-called higher-level preferences. Humans alone (see Povinelli & Bering, 2002; Povinelli & Giambrone, 2001) appear to be able to represent a model not only of the actual preference structure currently acted upon, but in addition a model of an idealised preference structure.

In this essay, I discuss two mental capacities that enable thinking about higher-level preferences. The first, cognitive decoupling, is an algorithmic-level construct (see Anderson, 1990; Dennett, 1987; Newell, 1982; Stanovich, 1999; Sterelny, 1990) and it will be discussed briefly because it has received considerable attention in disparate literatures in cognitive science. The second construct—a motive at the intentional level of mental understanding (e.g., Dennett, 1987; Newell, 1982)—is more of a thinking disposition and will be introduced in this essay.

COGNITIVE DECOUPLING ENABLES METAREPRESENTATION

In order to reason hypothetically, a person must be able to represent a belief as separate from the world it is representing. Numerous cognitive scientists have discussed so-called decoupling skills—the mental abilities that allow us to mark a belief as a hypothetical state of the world rather than a real one (e.g., Carruthers, 2002; Cosmides & Tooby, 2000; Dienes & Perner, 1999; Evans & Over, 1999, 2004; Frankish, 2004; Jackendoff, 1996; Sperber, 2000). Decoupling skills prevent our representations of the real world from becoming confused with representations of imaginary situations (simulations) that we create on a temporary basis in order predict the effects of future actions or to think of the consequences of pursuing alternative goals and desires. For example, when considering an alternative goal state different from the current goal state, one needs to be able to represent both—to represent one state of affairs as actual and another as hypothetical.

In short, a decoupled secondary representation—to use Perner's (1991) term—is necessary in order to avoid so-called representational abuse (Leslie, 1987)—the possibility of confusing our simulations with our

primary representations of the world as it actually is. The cognitive operation of decoupling, or what Nichols and Stich (2003) term cognitive quarantine, prevents our representations of the real world from becoming confused with representations of imaginary situations. In dual-process models, cognitive decoupling (outside the domain of behavioural prediction—so-called "theory of mind") is largely a System 2 operation. I have conjectured (Stanovich, 2004, in press) that the raw ability to sustain mental simulations while keeping the relevant representations decoupled is likely to be the key aspect of the brain's computational power that is being assessed by measures of fluid intelligence (on fluid intelligence, see Carroll, 1993; Horn & Cattell, 1967; Horn & Noll, 1997; Kane & Engle, 2002; Unsworth & Engle, 2005). Indeed, decoupling is the key operation of System 2 that accounts for its seriality and most importantly its computational expense. Such a view is consistent with much recent work on executive function and working memory (Conway, Kane, & Engle, 2003; Duncan, Emslie, Williams, Johnson, & Freer, 1996; Engle, 2002; Gray, Chabris, & Braver, 2003; Kane & Engle, 2002, 2003; Salthouse, Atkinson, & Berish, 2003; Unsworth & Engle, 2005, 2007).

Cognitive decoupling underlies the *ability* to reason about alternative preferences and to form higher-order preferences. However, it is not in itself the motive for forming higher-level preferences. There must be a motivational mechanism that creates the need for such self-evaluation. In the remainder of this essay I will speculate about the nature of this motive, and argue that it is a dispositional variable at a high level of generality that is, nonetheless, potentially measurable. I shall begin with a brief introduction to higher-order preferences, as they have been treated by philosophers and various decision scientists, including a discussion of Nozick's (1993) notion of rational integration: that we should strive for consistency in our preference hierarchy.

HIGHER-ORDER PREFERENCES AND RATIONAL INTEGRATION

There is a philosophical literature on the notion of higher-order evaluation of desires (Bratman, 2003; Dworkin, 1988; Harman, 1993; Lehrer, 1990, 1997; Lewis, 1989; Maher, 1993; Taylor, 1989; Watson, 1975), and it is one that is of potential theoretical interest for decision scientists (see Flanagan, 1996, for an insightful discussion that is informed by cognitive science). For example, in a classic paper on second-order desires, Frankfurt (1971) speculated that only humans have such metarepresentational 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 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. Wantons want—but they do not *care* what they want.

What I have been calling a critique of one's own desire structure can be more formally explicated in terminology more commonly used by economists, decision theorists, and cognitive psychologists (see Jeffrey, 1974; Kahneman & Tversky, 2000; Slovic, 1995; Tversky, Slovic, & Kahneman, 1990)—that is, in terms of second-order preferences (a preference for a particular set of first-order preferences). For example, imagine that: John prefers to smoke. Then using the preference relationship that is the basis for the formal axiomatisation of utility theory, we have:

S pref ~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, in preference notation:

(~S pref S) pref (S pref ~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. 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, pp. 243–247). This is because the struggle to achieve rational integration can destabilise first-order preferences in ways that make them more prone to the context effects that lead to the violation of the basic axioms of utility theory.

The struggle for rational integration is also what contributes to the feeling of alienation that people in the modern world often feel when contemplating the choices that they have made. People easily detect when their high-order preferences conflict with the choices actually made.

There is of course 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:
> [(~S pref S) pref (S pref ~S)] pref [S pref ~S]

We might in this case say that John's third-order judgement has ratified his second-order 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, staying out of smoky bars, etc.). On the other hand, a third-order judgement might undermine the second-order preference by failing to ratify it:

> John might prefer to smoke more than
> he prefers his preference to prefer not to smoke
> [S pref ~S] pref [(~S pref S) pref (S pref ~S)]

In this case, although John wishes he did not want to smoke, the preference for this preference is not as strong as his preference for smoking itself. We might suspect that this third-order judgement might not only prevent John from taking strong behavioural steps to rid himself of his addiction, but that over time it might erode his conviction in his second-order preference itself, thus bringing rational integration to all three levels.

Typically, philosophers have tended to bias their analyses towards the highest-level desire that is constructed—privileging the highest point in the regress of higher-order evaluations, using that as the foundation, and defining it as the true self. Modern cognitive science would suggest instead a Neurathian project in which no level of analysis is uniquely privileged. Philosopher Otto Neurath (1932/33; see Quine, 1960, pp. 3–4) employed the metaphor of a boat having some rotten planks. The best way to repair the planks would be to bring the boat ashore, stand on firm ground, and replace the planks. But what if the boat could not be brought ashore? Actually, the boat could still be repaired, but at some risk. We could repair the planks at sea by standing on some of the planks while repairing others. The project could work—we could repair the boat without being on the firm foundation of ground.

The Neurathian project is not guaranteed, however, because we might choose to stand on a rotten plank. Nothing in Frankfurt's (1971) notion of higher-order desires guarantees against higher-order judgements being infected by memes that are personally damaging (see Blackmore, 1999, 2005;

Dawkins, 1993; Dennett, 1991, 2006; Distin, 2005; Laland & Brown, 2002). For example, in the case of John the smoker, imagine that the two higher-order judgements had enough cognitive weight to lead him to take behavioural steps (therapy, etc.) to overturn his first-order desire, and he did so. It is, however, unnerving to realise that John's desire structure and resolution has exactly the same structure as that of the terrorist hijackers who destroyed thousands of lives in the September 11, World Trade Center attack.

An (obviously oversimplified) model might be that, like most people, even the hijackers—at least at one time—had a wanton desire for life over any religious martyrhood that they imagined. They preferred life to martyrhood:

LIFE pref MARTYR

But at some point in their lives, a faith-based memeplex found them to be good hosts and became the basis of a second-order judgement on their first-order desires. At some point in their lives, although they preferred life to martyrdom, they began to wish that they did not. They began to appreciate people who were terrorist martyrs even though they were not prepared to be one themself. They began to wish that they could be like those people. They preferred to prefer martyrhood to life:

(MARTYR pref LIFE) pref (LIFE pref MARTYR)

Perhaps the mismatched preference structure created the same discomfort that John, the smoker, felt after he had decided that he wished he did not prefer to smoke. The discomfort of the mismatched preference structure creates motivation to achieve rational integration. This might have spawned a third-order evaluation of the second-order preference—the person might have asked himself: am I right to have this second-order preference to prefer to prefer martyrdom? But since the terrorist is immersed in the same conceptual community that caused the original second-order judgement to be made—the same community that was the environment for the memeplex that caused the lack of rational integration in the first place—then it is likely that, as in the smoking example, the third-order judgement will ratify the second-order preference:

They will prefer their preference to prefer martyrhood to life over
their preference for life
[(MARTYR pref LIFE) pref (LIFE pref MARTYR)] pref [LIFE pref MARTYR]

But if that, as in the smoking example, begins the cognitive/behavioural cascade that leads to the flipping of the first-order preference, rational integration is achieved by murdering thousands of innocent people.

This unnerving example illustrates that there is no substitute for the Neurathian project of examining—in turn, and recursively—the planks represented by each level of judgement. We achieve coherence, float in our boat for a while, bring a different level of judgement into question, perhaps flip a preference relation causing incoherence that must be brought into reflective equilibrium again, this time perhaps by giving a different level of judgement priority.

Clearly we are not endorsing here a simple Platonic view where the lower mind (System 1) should always be ruled by the higher (System 2). There is a philosophical literature (e.g., Bennett, 1974; McIntyre, 1990) on cases like that of the Mark Twain character Huckleberry Finn. Huck helped his slave friend Jim run away because of very basic feelings of friendship and sympathy. However, Huck begins to have doubts about his action once he starts explicitly reasoning about how it is morally wrong for slaves to run away and for whites to help them. In this case, we want Huck to identify with the emotions emanating from System 1 modules and to reject the explicit morality that he has been taught.

The point of the terrorist and Huck Finn examples is that rational integration is not always achieved by simply flipping preferences that are in the minority across the levels of analysis; nor can it always best be achieved by the simple rule of giving priority to the highest level. Part of rational integration is the evaluation of the memes that form the values that are the basis of the higher-order evaluations. Note also that it is not just reflective higher-order preferences that are potentially meme-based. Certain higher-level evaluations can become so practised that they become reflexive, first-order goals. This of course is the purpose of the repetition of things like Victorian moral codes. Their promoters are hoping to make these judgements into reflexive responses rather than objects of analytic reflection (where processes of criticism might reject them).

Philosophers have thought that unless we had a level of cognitive analysis (preferably the highest one) that was foundational, something that we value about ourselves (various candidates in the philosophical literature have been personhood, autonomy, identity, and free will) would be put in jeopardy. Hurley (1989), in contrast, endorses a Neurathian view in which there does not exist either a "highest platform" or a so-called true-self, outside the interlocking nexus of desires. She argues that "the exercise of autonomy involves depending on certain of our values as a basis for criticizing and revising others, but not detachment from all of them, and that autonomy does not depend on a regress into higher and higher order attitudes, but on the first step" (p. 364). In short, the uniquely human project of self-definition begins at the first step, when an individual begins to climb the ladder of hierarchical values—when a person has, for the first time, a problem of rational integration.

But how do we know that a person is engaging deeply in such a process of self-definition and rational integration? Interestingly, perhaps the best indicator is when we detect a mismatch between a person's first-order and second-order desires with which the person is struggling: The person avows a certain set of values that imply that they should prefer to do something other than they do. However, when we do not detect such a struggle, the state of a person's preference hierarchy is ambiguous. The person might never engage in forming higher-order preferences—or they might well have a second-order preference that endorses their first-order preference and thus no internal struggle will be apparent in behaviour because there is none.

WHAT MOTIVATES THE SEARCH FOR RATIONAL INTEGRATION?

However, even if preference structures display inconsistencies, the mere existence of such inconsistency in and of itself is not enough to initiate the process of conflict resolution. No dual-process theorist would posit that rational integration is an automatic process. It is more likely an effortful System 2 process that needs motivational force to initiate. An important philosophical paper by Velleman (1992) provides guidance to psychologists in modelling and measuring the mechanisms involved in the process of rational integration. Velleman (1992) attempts to unpack Frankfurt's (1971) idea that "the agent's role is to adjudicate conflicts of motives" (p. 476). Under Frankfurt's view, Velleman points out (1992, p. 477), the role of adjudicator cannot be taken up by a higher-order attitude itself, because these will be always subject to review themselves:

> The functional role of agent is that of a single party prepared to reflect on, and take sides with, potential determinants of behavior at any level in the hierarchy of attitudes; and this party cannot be identical with any of the items on which it must be prepared to reflect or with which it must be prepared to take sides. Thus, the agent's role cannot be played by any mental states or events whose behavioral influence might come up for review in practical thought at any level.

Velleman then asks: "What mental event or state might play this role of always directing but never undergoing such scrutiny?" and answers (p. 477):

> It can only be a motive that drives practical thought itself. That is, there must be a motive that drives the agent's critical reflection on, and endorsement or rejection of, the potential determinants of his behavior, always doing so from a position of independence from the objects of review. Only such a motive would occupy the agent's functional role, and only its contribution to his behavior would constitute his own contribution.

In short, the concerns that drive practical thought in the search for rational integration cannot be Frankfurt's second-order desires themselves because these serve as the objects of reflection. Instead it must be something closer to the motive towards rational integration itself. So Velleman (p. 478) argues that

> some motive must be behind the processes of practical thought—from the initial reflection on motives, to the eventual taking of sides; and from second-order reflection to reflection at any higher level—since only something that was always behind such processes would play the causal role ordinarily attributed to the agent.

Velleman (1992) asks us to get used to that idea that practical thought itself (Manktelow, 2004; Millgram, 2001; Over, 2004) is propelled by a distinctive motive. For Velleman that motive is: "your desire to act in accordance with reasons, a desire that produces behavior, in your name, by adding its motivational force to that of whichever motives appear to provide the strongest reasons for acting" (p. 479). This attitude, for Velleman, performs the function of agency.

I would contend that what Velleman (1992) has identified is a motive at such a high level of generality that we might want to call it the Master Rationality Motive (MRM). It is the motive that drives the search for rational integration across our preference hierarchies. Importantly, Nozick (1993) has argued that it is not a particular algorithm for rational integration that is rational. Instead, what is rational is the *felt need* for rational integration. The need for rational integration is probably a function of the strength of the Master Rationality Motive, and individual differences in the former probably arise because of individual differences in the latter. Thus, the Master Rationality Motive (MRM) is what sustains the search for rational integration.

THE MASTER RATIONALITY MOTIVE AS A PSYCHOLOGICAL CONSTRUCT

The psychological literature on individual differences in rational thought does contain some attempts to measure the MRM. Epstein's Head Over Heart scale (a precursor to his rational-experiential inventory; see Epstein, Pacini, Denes-Raj, & Heier, 1996; Epstein, Pacini, Heier, & Denes-Raj, 1995; Pacini & Epstein, 1999) is perhaps the measure with the most overlap with the MRM (for relevant empirical work, see Bartels, 2006; Klaczynski & Lavallee, 2005; Newstead, Handley, Harley, Wright, & Farrelly, 2004). Our own actively openminded thinking scale (Stanovich & West, 2007) taps a partially overlapping construct. Recently, however, we have tried to be more systematic, and have compiled items from a variety of existing scales and have tested new ones in an attempt to measure the Master Rationality

Motive in questionnaire form. A proposed MRM scale is presented in Table 1.

My conjecture is that the MRM scale will tap an individual difference characteristic different from cognitive ability (intelligence). Thus, it is proposed that it is a scale that could capture unique variance in rational thinking after cognitive ability has been partialled out. One reason for making this conjecture is that there are already empirical indications that more micro thinking dispositions can predict variance in reasoning tasks. For example, various thinking dispositions such as need for cognition and actively openminded thinking have been found to predict (sometimes after control for cognitive ability) performance on a host of tasks from the heuristics and biases literature (Bartels, 2006; Chatterjee, Heath, Milberg, & France, 2000; Klaczynski & Lavallee, 2005; Kokis, Macpherson, Toplak, West, & Stanovich, 2002; McElroy & Seta, 2003; Newstead et al., 2004; Pacini & Epstein, 1999; Parker & Fischhoff, 2005; Perkins & Ritchhart,

TABLE 1
Items on the Master Rationality Motive scale

Item	*Source*
Intuition is the best guide in making decisions. (R)	Stanovich & West (1997)
Certain beliefs are just too important to abandon no matter how good a case can be made against them. (R)	Sá, West, & Stanovich (1999)
I am only confident of decisions that are made after careful analysis of all available information.	Leary, Shepperd, McNeil, Jenkins, & Barnes, (1986)
I do not like to be too objective in the way I look at things. (R)	Leary et al. (1986)
After I make a decision, it is often difficult for me to give logical reasons for it. (R)	Leary et al. (1986)
I believe in following my heart more than my head. (R)	Epstein et al. (1995)
It is more important to me than to most people to behave in a logical way.	Epstein et al. (1995)
I like to gather many different types of evidence before I decide what to do.	New item
I don't feel I have to have reasons for what I do. (R)	New item
I like to think that my actions are motivated by sound reasons.	New item
I like to have reasons for what I do.	New item
I don't like to have to justify my actions. (R)	New item
If a belief suits me and I am comfortable, it really doesn't matter if the belief is true. (R)	New item
Item #2 from Facet 6 (Deliberation) of the Conscientious domain of the NEO PI-R	Costa & McCrae (1992)
Item #4 from Facet 6 (Deliberation) of the Conscientious domain of the NEO PI-R	Costa & McCrae (1992)

(R) = item reversed scored.

2004; Shiloh, Salton, & Sharabi, 2002; Simon, Fagley, & Halleran, 2004; Smith & Levin, 1996; Stanovich & West, 1998, 1999, 2000; Toplak & Stanovich, 2002; Verplanken, 1993).

However, this previous literature differs from what I am proposing here in two ways. First, the thinking dispositions that have been studied are at a lower level of generality than the MRM (Perkins & Ritchhart, 2004). Second, the criterion variables have been limited to aspects of instrumental rationality—for example, how well people satisfy the choice axioms of utility theory or the strictures of Bayesian belief updating. As such, the criterion variables have reflected aspects of what has been termed a thin theory of rationality (see Elster, 1983). Thin theories define rationality in terms of current beliefs and desires. However, it is a broad notion of rationality that is most likely to have a unique connection to the MRM.

Broad rationality encompasses a cognitive critique of the beliefs and desires that are input into the implicit calculations that result in instrumental rationality (Elster, 1983). Moving to a broad theory of rationality will necessitate the evaluation of the content of desires and goals. Although the axioms of instrumental rationality are well articulated, the criteria for evaluating broad rationality are much more complex and contentious. Nonetheless, some attempts have been made. For example, it has been argued by several theorists that desires that, upon reflection, we would rather eliminate than fulfil are irrational (see Nathanson, 1994). Other theorists argue that conflicting desires, or desires based on false beliefs, are irrational. Finally, it could be argued that the persistent tendency to take actions whose expected utility is different from their experienced utility is a sign of irrationality (Kahneman, 1994, 1999, 2003; Kahneman & Snell, 1990; Kahneman, Wakker, & Sarin, 1997).

Nozick (1993) has made a somewhat more formal attempt to articulate the principles for the evaluation of the rationality of preferences (see his discussion of 23 criteria for the evaluation of preferences, especially principles IV, VII, X, and XIV). Examples of his principles include: the degree to which a person finds lack of rational integration aversive and is willing to take steps to rectify it; whether the individual can state a reason for all second-order desires; whether it is the case that a person's desires are not such that acting on them leads to irrational beliefs; whether a person avoids forming desires that are impossible to fulfil; the degree of higher-level evaluation undertaken; and others (see Nozick, 1993).

A broad theory of rationality will also emphasise the interplay between epistemic and practical rationality in some interesting ways. For example, it is relatively uncontroversial that it is rational to have derived goals. One might desire to acquire an education not because one has an intrinsic desire for education, but because one desires to be a lawyer and getting an education leads to the fulfilment of that goal. Once derived goals are

allowed, the possibility of evaluating goals in terms of their comprehensiveness immediately suggests itself. That is, we might evaluate certain goals positively because attaining them leads to the satisfaction of a wide variety of *other* desires. In contrast, there are some goals whose satisfaction does not lead to the satisfaction of any other desires. In the extreme there may exist goals that are in conflict with other goals. Fulfilling such a goal would actually impede the fulfilment of other goals.

In the vast majority of mundane cases, one derived goal that would serve to fulfil a host of others is the goal of wanting one's beliefs to be true. Obviously, perfect accuracy is not required and, equally obviously, there is a point of diminishing returns where additional cognitive effort spent in acquiring true beliefs will not have adequate payoff in terms of goal achievement. Nevertheless, other things being equal, the presence of the desire to have true beliefs will have the long-term effect of facilitating the achievement of a host of goals. It is a superordinate goal in a sense and—again, except for certain bizarre cases—should be in the desire network of most individuals. Thus, even if epistemic rationality really is subordinate to practical rationality, the derived goal of maintaining true beliefs will mean that epistemic norms must be adhered to in order for practical goals to be achieved.

My purpose here is not to provide a complete discussion of broad rationality in the domain of desires and goals, but instead to simply sketch the type of motive that might initiate the attempt to achieve rational integration between first-order and second-order preferences. This essay itself was in part meant to provoke a psychological research programme that encompasses empirical investigations of the degree to which people engage in a cognitive critique of their beliefs and desires. It is conjectured here that the Master Rationality Motive will be strongly connected to differences in the extent of this cognitive critique.

Of course I do not wish to paint the MRM in too essentialist a fashion before investigation has even begun. The extent to which it is a generic motive is yet to be determined. There may well be a good degree of domain specificity in the extent to which people strive to achieve rational integration. Nonetheless, some degree of domain specificity might not necessarily be antithetical to the MRM construct. The logic of memetics has suggested that there are individual differences not only among people but also among the beliefs that they host. Importantly, beliefs differ in the extent of their so-called adversarial properties (Blackmore, 1999; Dawkins, 1993; Lynch, 1996)—how strongly they are structured to repel competing ideas. Thus, how much the beliefs that are the source of higher-order evaluations encourage or repel the process of rational integration may be quite variable. This implies that domain specificity in rational integration could well exist without domain specificity in the MRM construct itself. Issues such as this

are the type of question about the MRM that future research will need to answer.

Human thinking is not just confined to the pursuit of instrumental rationality. Humans are also concerned about issues of broad rationality. They attempt to critique current desires and goals using their metarepresentational abilities. It is conjectured here that the critique itself—and the consistency issues it raises—is driven by a distinctive thinking disposition at a high level of generality, the Master Rationality Motive. Preliminary research on related constructs encourages the view that this construct is measurable.

Manuscript received 30 October 2006
Revised manuscript received 3 April 2007
First published online 28 June 2007

REFERENCES

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bartels, D. M. (2006). Proportion dominance: The generality and variability of favoring relative savings over absolute savings. *Organizational Behavior and Human Decision Processes*, *100*, 76–95.

Bennett, J. (1974). The conscience of Huckleberry Finn. *Philosophy*, *49*, 123–134.

Blackmore, S. (1999). *The meme machine*. New York: Oxford University Press.

Blackmore, S. (2005). Can memes meet the challenge? In S. Hurley & N. Chater (Eds.), *Perspectives on imitation* (Vol. 2, pp. 409–411). Cambridge, MA: MIT Press.

Bratman, M. E. (2003). Autonomy and heirarchy. *Social Philosophy & Policy*, *20*(2), 156–176.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

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

Chatterjee, S., Heath, T. B., Milberg, S. J., & France, K. R. (2000). The differential processing price in gains and losses: The effects of frame and need for cognition. *Journal of Behavioral Decision Making*, *13*, 61–75.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Science*, *7*, 547–552.

Cosmides, L., & Tooby, J. (2000). Consider the source: The evolution of adaptations for decoupling and metarepresentation. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 53–115). Oxford, UK: Oxford University Press.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory*. Odessa, FL: Psychological Assessment Resources.

Dawkins, R. (1993). Viruses of the mind. In B. Dahlbom (Ed.), *Dennett and his critics* (pp. 13–27). Cambridge, MA: Blackwell.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little Brown.

Dennett, D. C. (2006). From typo to thinko: When evolution graduated to semantic norms. In S. C. Levinson & P. Jaisson (Eds.), *Evolution and culture* (pp. 133–145). Cambridge, MA: MIT Press.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*, 735–808.

Distin, K. (2005). *The selfish meme*. Cambridge, UK: Cambridge University Press.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, *30*, 257–303.

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge, UK: Cambridge University Press.

Elster, J. (1983). *Sour grapes: Studies in the subversion of rationality*. Cambridge, UK: Cambridge University Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390–405.

Epstein, S., Pacini, R., Heier, H., & Denes-Raj, V. (1995). *Individual differences in intuitive and analytical information processing*. Unpublished manuscript.

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. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, *13*, 378–395.

Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. New York: Psychology Press.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology 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.

Flanagan, O. (1996). *Self expressions: Mind, morals, and the meaning of life*. New York: Oxford University Press.

Flanagan, O. (2002). *The problem of the soul*. New York: Basic Books.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, *68*, 5–20.

Frankish, K. (2004). *Mind and supermind*. Cambridge, UK: Cambridge University Press.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316–322.

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.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 1–23.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53–91). New York: Guilford Press.

Hurley, S. L. (1989). *Natural reasons: Personality and polity*. New York: Oxford University Press.

Jackendoff, R. (1996). How language helps us think. *Pragmatics and Cognition*, *4*, 1–34.

Jeffrey, R. (1974). Preferences among preferences. *Journal of Philosophy*, *71*, 377–391.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 3–25). Thousand Oaks, CA: Sage.

Kahneman, D. (2003). Experienced utility and objective happiness: A moment-based approach. In I. Brocas & J. D. Carrillo (Eds.), *The psychology of economic decisions: Vol. 1. Rationality and well-being* (pp. 187–208). Oxford, UK: Oxford University Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgement* (pp. 49–81). New York: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). New York: Cambridge University Press.

Kahneman, D., & Snell, J. (1990). Predicting utility. In R. M. Hogarth (Ed.), *Insights into decision making* (pp. 295–310). Chicago: University of Chicago Press.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, UK: Cambridge University Press.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *The Quarterly Journal of Economics*, *112*(2), 375–405.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review*, *9*, 637–671.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Klaczynski, P. A., & Lavallee, K. L. (2005). Domain-specific identity, epistemic regulation, and intellectual ability as predictors of belief-based reasoning: A dual-process perspective. *Journal of Experimental Child Psychology*, *92*, 1–24.

Kokis, J., Macpherson, R., Toplak, M., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*, 26–52.

Laland, K. N., & Brown, G. R. (2002). *Sense and nonsense: Evolutionary perspectives on human behaviour*. Oxford, UK: Oxford University Press.

Leary, M. R., Shepperd, J. A., McNeil, M. S., Jenkins, B., & Barnes, B. D. (1986). Objectivism in information utilization: Theory and measurement. *Journal of Personality Assessment*, *50*, 32–43.

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.

Leslie, A. M. (1987). Pretense and representation: The origins of "Theory of Mind". *Psychological Review*, *94*, 412–426.

Lewis, D. (1989). Dispositional theories of value. *Proceedings of the Aristotelian Society, Supplementary Volume 63*, 113–137.

Lynch, A. (1996). *Thought contagion*. New York: Basic Books.

Maher, P. (1993). *Betting on theories*. Cambridge, UK: Cambridge University Press.

Manktelow, K. I. (2004). Reasoning and rationality: The pure and the practical. In K. I. Manktelow & M. C. Chung (Eds.), *Psychology of reasoning: Theoretical and historical perspectives* (pp. 157–177). Hove, UK: Psychology Press.

McElroy, T., & Seta, J. J. (2003). Framing effects: An analytic-holistic perspective. *Journal of Experimental Social Psychology*, *39*, 610–617.

McIntyre, A. (1990). Is akratic action always irrational? In O. Flanagan & A. O. Rorty (Eds.), *Identity, character, and morality* (pp. 379–400). Cambridge, MA: MIT Press.

Millgram, E. (Ed.). (2001). *Varieties of practical inference*. Cambridge, MA: MIT Press.

Nathanson, S. (1994). *The ideal of rationality*. Chicago: Open Court.

Neurath, O. (1932/33). Protokollsatze. *Erkenntis*, *3*, 204–214.

Newell, A. (1982). The knowledge level. *Artificial Intelligence*, *18*, 87–127.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., & Farrelly, D. (2004). Individual differences in deductive reasoning. *Quarterly Journal of Experimental Psychology*, *57A*, 33–60.

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. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgement and decision making* (pp. 3–18). Malden, MA: Blackwell Publishing.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual differences approach. *Journal of Behavioral Decision Making*, *18*, 1–27.

Perkins, D., & Ritchhart, R. (2004). When is good thinking? In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 351–384). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.

Povinelli, D. J., & Bering, J. M. (2002). The mentality of apes revisited. *Current Directions in Psychological Science*, *11*, 115–119.

Povinelli, D. J., & Giambrone, S. (2001). Reasoning about beliefs: A human specialization? *Child Development*, *72*, 691–695.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Sá, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*, 497–510.

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General*, *132*, 566–594.

Shiloh, S., Salton, E., & Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, *32*, 415–429.

Simon, A. F., Fagley, N. S., & Halleran, J. G. (2004). Decision framing: Moderating effects of individual differences and cognitive processing. *Journal of Behavioral Decision Making*, *17*, 77–93.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

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

Smith, S. M., & Levin, I. P. (1996). Need for cognition and choice framing effects. *Journal of Behavioral Decision Making*, *9*, 283–290.

Sperber, D. (2000). Metarepresentations in evolutionary perspective. In D. Sperber (Ed.), *Metarepresentations: 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. (in press). *Rationality and the tri-process theory of mind*. New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*, 342–357.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.

Stanovich, K. E., & West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. *Cognitive Psychology*, *38*, 349–385.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–726.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*, 225–247.

Sterelny, K. (1990). *The representational theory of mind: An introduction*. Oxford, UK: Basil Blackwell.

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

Toplak, M., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *94*, 197–209.

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

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence*, *33*, 67–81.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132.

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

Verplanken, B. (1993). Need for cognition and external information search: Responses to time pressure during decision-making. *Journal of Research in Personality*, *27*, 238–252.

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