The Locality and Globality of Instrumental Rationality: The normative significance of preference reversals

Brian Kim*

Abstract

Abstract When we ask a decision maker to express her preferences, it is typically assumed that we are eliciting a pre-existing set of preferences. However, empirical research has suggested that our preferences are often constructed on the fly for the decision problem at hand. This paper explores the ramifications of this empirical research for our understanding of instrumental rationality. First, I argue that these results pose serious challenges for the traditional decision-theoretic view of instrumental rationality, which demands global coherence amongst all of one's beliefs and desires. To address these challenges, I first develop a minimal notion of instrumental rationality that issues in localized, goal-relative demands of coherence. This minimal conception of instrumental rationality is then used to offer a more sophisticated account of the global aspects of instrumental rationality. The resulting view abandons all-or-nothing assessments of rationality and allows us to evaluate decision makers as being rational to varying degrees. My aim is to propose a theory that is both psychologically and normatively plausible.

Decision theory offers a familiar view of deliberation and instrumental rationality. It depicts human agents as possessing a set of standing beliefs and desires that are expressed when they make choices and act. For when an agent deliberates, her beliefs and desires are used to evaluate the available choices, and she acts in the way that she believes will result in the most desirable consequences. Instrumental rationality then demands that decision makers (henceforth DMs) possess a coherent set of preferences for any given deliberative context. And since these preferences are determined by beliefs and desires, instrumental rationality demands that DMs possess a coherent set of beliefs and desires, which means that these beliefs and desires ought to be representable respectively as probabilities and utilities.[1] Let's call this decision-theoretic demand, **Deliberative Coherence**.

To fully articulate the nature of this demand, we must survey some familiar and unfamiliar details of the decision-theoretic view of deliberation. What is familiar is that DMs face decision problems that are defined by a set of *acts* that the DM is choosing between along with the *consequences* that these acts will produce when one of a mutually exclusive and exhaustive set of *states* obtains. Consequences are propositions that describe what might result from the DM's actions. States are propositions that describe the ways the world might be independent of the DM's actions. The DM's preferences between acts are then determined by her beliefs and desires about, respectively, the states and consequences.[2]

*For their thoughtful comments and objections, I would like to thank Achille Varzi, Anubav Vasudevan, Dana Howard, Don Hubin, Glenn Ross, Guillermo Del Pinal, John Brunero, John Collins, Katie Gasdaglis, Mark Alfano, Sigrun Svavasdottir, and two anonymous referees for Synthese. I also thank the audiences at Kansas State, Columbia University, the Decisions, Games & Logic Workshop, and the Canadian Society for Epistemology for their comments on earlier versions of the paper.

[1] While it will not affect my discussion, I will for the sake of concreteness take for granted the standard set of structural requirements on preferences articulated in (Savage 1972). See Part I of (Anand, Pattanaik, and Puppe 2009) for a survey of alternative requirements. I will also be adopting the realistic as opposed to formalistic interpretation of decision theory on which an expected utility function is a measure of beliefs and desires rather than an indicator function that offers a "definitional reformulation" of the DM's preferences (See pp. 144-146 in Hansson 1988). Of course, the DM may possess "brute" preferences that are inconsistent with those entailed by her beliefs and desires. However, to avoid confusion between these two ways of talking about preferences, I will avoid talk of such brute preferences.

[2] I will be assuming the separability of belief and desire. See (Levi 1999) and (Jeffrey 1965) for opposing sides of this debate.

Before we articulate a crucial yet unfamiliar detail, we must first make an important conceptual distinction between:

Deliberative judgments: beliefs about states and desires about consequences that determine a DM's preferences over the acts considered in a specific decision problem[3]

***Background attitudes*:** beliefs and desires that are fairly stable from one decision problem to the next

Deliberative judgments are beliefs- and desires-[in a decision problem]. Call these *deliberative beliefs and desires*. In contrast, background attitudes are beliefs and desires simpliciter. So we could, for example, distinguish desiring-[when choosing between vacations] health more than comfort from desiring health more than comfort simpliciter.

Given this distinction, Deliberative Coherence is strictly speaking the demand that a DM's deliberative beliefs and desires determine a coherent set of preferences. However, decision theorists typically assume that Deliberative Coherence has broader normative significance, which can be articulated by the following two theses:

Completeness of Deliberative Coherence (COMPLETE): A DM is instrumentally rational if and only if she satisfies Deliberative Coherence.[4]

Extended Scope of Deliberative Coherence (EXTENDED): The scope of Deliberative Coherence extends over a DM's background attitudes.[5]

COMPLETE expresses the commitment to the Humean view of rationality embraced by many decision theorists.[6] EXTENDED, in contrast, is rarely made explicit but becomes obvious once we register the distinction between deliberative judgments and background attitudes. This unspoken commitment seems obvious and justified by a seemingly innocuous empirical assumption that deliberative judgments are direct expressions of a DM's background attitudes. After all, if one believes-*[when deciding what to wear]* that it will rain, it is assumed that one believes that it will rain. Belief and desire are not typically indexed to a particular decision problem. Thus, coherence constraints that govern deliberative judgments are ipso facto coherence constraints that govern background attitudes. So the demands for coherence extend from the judgments used in particular decision problems to the background attitudes that these judgments express.

The aim of this paper is two-fold. I will first argue that the standard decision-theoretic account of instrumental rationality, as articulated by COMPLETE and EXTENDED, must be abandoned. I will then develop an alternative decision-theoretic account of instrumental rationality by reconsidering the broader normative significance of Deliberative Coherence. In §1, I summarize empirical results that raise important challenges for the standard decision-theoretic account. In §2, I consider responses to these challenges and conclude against COMPLETE and EXTENDED that Deliberative Coherence does not provide a complete account of rationality and should be understood as issuing only in a local demand for coherence governing judgments within a particular decision problem. In §3, I appeal to, what I call, a local and context-sensitive interpretation of Deliberative Coherence to account for more global features of instrumental rationality thereby developing an alternative to the standard decision-theoretic account. The resulting account allows for more nuanced and fine-grained evaluations of rationality, and we will be able to to account for the fact that we often judge the extent which DM's are rational

[3] I use the term 'judgment' only to highlight that these attitudes are *used* to deliberate. I am not assuming that they are conscious, under the agent's control, or truth-evaluable in the case of desires.

[4] "The overarching goal of normative decision theory is to establish a general standard of rationality for the sort of instrumental (or "practical") reasoning that people employ when trying to choose means appropriate for achieving ends they desire." (Joyce 1999: 9)

[5] Since it is typically assumed that one's standing mental states ought to be coherent, this thesis is implicitly accepted in discussions of rational coherence. For example, Ralph Wedgwood writes, "Instrumental irrationality crucially consists in this combination of mental states." (Wedgwood 2011: 290) No explicit reference to the deliberative context is made.

[6] For example, Maurice Allais writes, "it cannot be too strongly emphasized *that there are no criteria for the rationality of ends as such other than the condition of consistency.* Ends are completely arbitrary." (Allais 1979: 70)

and not merely whether or not the DM is rational. By accommodating the relevant empirical data and addressing some foundational normative problems, my aim is to present a theory that is both psychologically and normatively plausible.

1 Preference Reversals

In the 20th century, most systematic theories of human choice behavior (e.g., classical economics and behavioral decision theory) assumed that human decision makers have context-independent preferences.[7] On this view, the preferences revealed through choice behavior do not depend upon the particular deliberative task that an agent faces. This assumption is tantamount to the claim that deliberative judgments directly express background attitudes. Beginning in the late 1960s, Paul Slovic and Sarah Lichtenstein challenged this assumption and showed that a subject's expressed preferences often varied depending upon the way that they were elicited.[8] The phenomenon was described as a *preference reversal* because subjects expressed preferences between pairs of choices that would then be reversed when the same preferences were elicited in a different manner.

To illustrate, consider the following task in which college students were presented with pairs of job offers.[9] They were asked to imagine that the offers were identical except with respect to two attributes, annual salary and vacation days. Their preferences were first elicited using the CHOICE task where subjects were asked to choose the offer that they preferred or express their indifference. For example, would you prefer *Job A: $31,500, 15 vacation days* or *Job B: $36,500, 5 vacation days*?

After some distractions – feel free to provide your own – the subjects' preferences were elicited using the MATCH task. Here, subjects were presented with both attributes for Job A but only provided with the number of vacation days for Job C. They were then asked to specify the annual salary for Job C such that they would be indifferent between the two options (Table 1). At what salary would you be indifferent?

	Salary	**Vacation Days**
Job A	$31,500	15 days
Job C	x	5 days

Table 1: MATCH Task

The missing value for Job C can be used to infer a subject's preferences between Job A and B.[10] If $x = \$36,500$, then the subject should be indifferent between A and B since B would be identical to C. If $x > \$36,500$, then A would be preferred to B since B weakly dominates C. For similar reasons, if $x < \$36,500$, then B would be preferred to A. As it turned out, the majority of subjects expressed a preference for B over A in CHOICE while specifying a value for x that entailed a preference for A over B in MATCH, thus reversing their preferences.

The phenomenon has been subjected to rigorous testing using a variety of choices and modes of elicitation in both artificial laboratory settings and real-life betting scenarios. The results have been

[7] "In conventional economic theory, preferences are taken as primitive. A fundamental assumption is that preferences are independent of the tasks that an agent faces... It is crucial for many standard methods of economic analysis and implicit both in traditional models, such as expected utility theory, and in generalisations of them." (Cubitt, Munro, and Starmer 2004: 709) (Hausman 2012) offers an in-depth argument that the preferences appealed to in formal economic theory should be understood as context-independent all-things-considered judgments.

[8] (Lichtenstein and Slovic 1968) and (Lichtenstein and Slovic 1971) present early findings.

[9] This study comes from (Fischer et al. 1999).

[10] This inference assumes that subjects do not prefer dominated choices. (Cubitt, Munro, and Starmer 2004) tested this assumption and found that violations of dominance could not explain preference reversals.

replicated across a wide range of circumstances.[11] Glenn Shafer concludes that "it is the most fundamental result of three decades of empirical investigation. The preferences people express are unstable. They depend on the questions asked."[12] Fischer *et al.* write that from the point of view of traditional economic theory, decision analysis, and other management science disciplines, it is "[assumed] that people have rational preferences"; in contrast, "behavioral decision research suggests an alternative view, according to which people construct preferences as they are needed."[13] On this **Constructive View of Deliberation**, the preferences that human DMs use to deliberate depend upon context-sensitive features of a decision problem.[14]

This constructive view abandons the account of deliberation that has implicitly justified COMPLETE. Traditionally, the decision theorist's job is to articulate coherence constraints that the DM can use to adjust her attitudes until they reach a rational equilibrium. The role of the normative theory is restricted to this minimal type of coherence because it is assumed that the DM has a fairly rich set of pre-existing, context-independent beliefs and desires that are used to deliberate. However, since our deliberative judgments are often not well-defined outside of a deliberative context, a normative theory that only articulates coherence constraints governing these judgments is incomplete. If a DM has not made any deliberative judgments or if the judgments she uses depend crucially on features of the deliberative context, then the principles of instrumental rationality should also govern this context-sensitive process of construction. So satisfying Deliberative Coherence would not be sufficient to meet the demands of instrumental rationality. Therefore, we should reject COMPLETE since it ignores the construction and context-sensitivity of preference.

The argument for EXTENDED also depends upon the assumption that deliberative judgments are direct expressions of background attitudes. However, according to the constructive view, the former are context-dependent while the latter are context-independent. Since this empirical assumption cannot be used to extend the scope of Deliberative Coherence, this raises a question: what exactly is its scope? Which attitudes and judgments ought to be coherent for a DM to count as instrumentally rational? Let us turn to consider these questions.

2 The Scope of Deliberative Coherence

Deliberative Coherence states that a DM's preferences for a particular decision problem ought to be coherent. However, a DM Steve may satisfy Deliberative Coherence even though his preferences when choosing between summer vacations do not cohere with his preferences when choosing between lunch options. For example, he may prefer seafood over steak in one context but not the other. EXTENDED naturally expands the scope of Deliberative Coherence, and I suggested that the implicit argument for EXTENDED relied upon the empirical assumption that Steve's preferences when planning summer vacations are determined by the same beliefs and desires that determine his preferences when planning lunch. However, it should be surprising for a normative theory to depend so crucially upon a questionable empirical assumption. So perhaps we can defend EXTENDED by replacing the empirical assumption with a normative thesis that articulates the scope of Deliberative Coherence in a way that would extend its demand to our background attitudes.

2.1 Global Coherence and All-Things-Considered Rationality

One intuitive proposal is to extend the local demand for coherence by considering the relationship between a DM's preferences across different decision problems. Intuitively, the preferences Steve uses

[11] See Part II of (Lichtenstein and Slovic 2006).

[12] (Shafer 1986: 464-465)

[13] (Fischer et al. 1999: 1057) Preference reversals are not the only evidence for the construction of preference. See Part IV of (Lichtenstein and Slovic 2006) for additional examples.

[14] This argument does not rest upon a specific view of rational choice. (Cubitt, Munro, and Starmer 2004) show that for certain types of preference reversals, the subject's expressed preferences across different modes of elicitation cannot be represented by any context-independent choice function.

to evaluate his vacation and lunch options ought to form a coherent whole. Suppose X and Y are dining options that respectively offer excellent seafood and excellent steak. If Steve has already expressed a preference for seafood while planning his vacation, then if X and Y are otherwise identical, it would seem that he ought to prefer X over Y. Moreover, this inter-contextual demand may arise from a certain view of how our decision problems fit together. All of our decision problems are really part of a single overarching decision problem of how to live, and if one is evaluating the possible choices (i.e., life strategies) coherently in that global decision problem, one's beliefs and desires in narrowly defined decision problems must be coherent with one another.

We can capture these intuitions with the following demand that makes explicit the larger scope of Deliberative Coherence.

Global Coherence: Across all decision problems, a DM's preferences should be coherent with one another.

Global Coherence is the synchronic demand that a DM ought to possess a single coherent set of preferences that can be used for any decision problem. So the deliberative beliefs and desires that determine preferences in different decision problems ought to be coherent with one another.

Of course, decision problems can be about anything. Consider the arbitrary consequence of owning a donkey wearing a blue hat. One could always come across a decision problem in which this consequence must be considered. For example, this donkey may be the prize of a lottery. For similar reasons, every possible state must also be considered. So if one must possess a single coherent way of evaluating choices across all possible decision problems, then Global Coherence entails the view that instrumental rationality is an:

All-Things-Considered Rationality: A DM is rational if and only if she has a coherent set of preferences that are determined by the DM's beliefs and desires about respectively every possible state and consequence.[15]

As we had hoped, the all-things-considered view of instrumental rationality endorses EXTENDED. For it requires DMs to possess a single coherent set of preferences that can then be used in each and every decision problem. This demand is equivalent to EXTENDED's demand that one possess a set of coherent background attitudes. Both simply amount to the synchronic demand that DMs possess a comprehensive set of coherent preferences.

Unfortunately, there is a serious problem with Global Coherence and all-things-considered rationality. They violate the maxim that 'ought' implies 'can' in a particularly heinous way. All-things-considered rationality requires that we consider every possible state and consequence. In order to consider every possible state, one must consider the set of mutually exclusive and exhaustive set of states that describes the world in all its detailed glory. This is called a set of *complete states*. A complete state offers an answer to every possible question the DM could raise about the way the world might be, from whether corn grows on the eastern shore of Lake Winnebago to where every particle in the universe is located. Similarly, complete consequences offer an answer to every conceivable question the DM could raise about what may happen to her. Complete acts then describe possible actions in terms of their completely described consequences given each completely described state. Global Coherence requires DMs to consider complete states and consequences but human agents are incapable of doing so. The propositions that pick these out are too complex to be grasped, to be the contents of human judgments. And if someone cannot grasp the propositions that pick out complete states and consequences, they cannot possess all-things-considered preferences. Thus, the demand that such preferences be coherent would never apply.[16]

[15] "Practical reason, it might be suggested, is a holistic enterprise, properly concerned not merely with identifying means to the realization of individual ends, but with the coordinated achievement of the totality of an agent's ends." To offer just a small sampling, (Wallace 2008), (Fantl and McGrath 2002), (Wedgwood 2007), (Gibbard 2009), and (Broome 2013) either endorse or appeal to an all-things-considered notion of rationality.

[16] One might object that DMs are capable of grasping complete states and consequences by doing so in a piecemeal fashion. For example, a DM may be capable of considering states that describe both the size and color of an object by

To clarify what I take the problem to be, the following is meant to be an analogous case. Suppose I am considering a number of sounds meant to startle dogs. I ask you to listen and rank the sounds in terms of how startling they might seem. In order to be coherent, I demand that your ranking be transitive. Now, the task has an important twist. The alarms only produce high-frequency sounds that have no effect on humans. In this case, any demand I make about your judgments seems silly. For in evaluating these sounds, the demand to have coherent judgments makes sense only if one can make some meaningful judgments about how startling they might be to dogs. And since this condition is not met, the demands that govern this task are not applicable to human agents. The reason is simple. Humans can't engage in the relevant task.

Similarly, since all-things-considered rationality requires that we consider states and consequences that are beyond our grasp, its demands are just as inapplicable. When deciding between complete acts (i.e. life strategies), the demand for Global Coherence makes sense only if one can make some meaningful all-things-considered judgments. Since human beings fail to meet this condition, all-things-considered rationality fails to supply applicable demands. We should understand the failure of applicability in the following way. All-things-considered rationality governs deliberative activities in which the DM considers complete states and consequences. So the demands of all-things-considered rationality do not apply to humans because they do not govern activities that are in the realm of human possibility.[17]

2.2 Adequate Coherence and All-Things-Relevant Rationality

Underlying Global Coherence is the intuition that one should be coherent across decision problems. However, this intercontextual demand for coherence is applicable – to supply norms that govern activities within the realm of human possibility – only if there is a way to restrict its scope. It might seem natural to restrict the scope of this inter-contextual demand for coherence by restricting the relevant set of decision problems. However, it's hard to see how to do this is a principled way. How would we demarcate the decision problems that must be accounted for? Which preferences ought to be coherent with one another?

While no principled demarcation seems viable, we could appeal to the vague demand that DMs do the best they can given their cognitive abilities and practical situation.[18] Depending on how we interpret this idea, the proposal is that we leave the scope of Deliberative Coherence as an open, vague, or context-sensitive parameter.[19] This parameter once fixed would demarcate the set of decision

independently considering answers to questions about the size and color of the object. However, this way of considering the details of a state or consequence is not sufficient for accomplishing the DM's deliberative task. Allow me to briefly summarize my response to this objection. For a more detailed version of the argument, see (Kim 2012). As the preference reversal phenomenon has shown, the preferences of human agents can vary depending upon how we ask questions. Thus, the DM's belief about a state may differ depending upon whether she independently considers the potential size and color of an object or considers the size and color together. In the case of belief, the reason for this instability is simple. The DM may believe that there is a correlation between the object's size and color and so her beliefs may vary having considered these two questions together rather than having considered these two questions independently. Similarly, the DM's desires may differ depending upon whether she considers answers to questions about what will happen to her independently or all at once. The reasons for the instability of desire come not from correlations between states of affairs but from the complex ways our values interact. For example, it is not obvious how to combine a ranking of jobs in terms of salary with a ranking of jobs in terms of location. To do so, one must consider at once the salary and location of a job, which means that one must be able to have thoughts whose contents are fine-grained enough to describe both salary and location. Therefore, to consider a complete state or consequence, one must be able to have thoughts whose contents are fine-grained enough to describe the answer to every possible question all at once. My claim here is that these contents are too fine-grained to be the contents of human thoughts.

[17]These criticisms and the positive proposal that follows may be situated within the bounded rationality research program first articulated by Herb Simon. "Broadly stated, the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist." (Simon 1955: 99)

[18]I thank an anonymous reviewer for this suggestion.

[19]As I have stated it, the demand seems questionable. Why should we be as coherent as humanly possible? What do we achieve by meeting this demand? In what sense are we doing better as we become coherent across more and more decision problems? And even if we are doing better, are we doing better in the relevant sense? Are we being more rational

problems that should be accounted for. While I will ultimately endorse a view of this kind, there are two reasons that we should, for the moment, set it aside. First, by embracing such a demand, we drastically alter the standard decision-theoretic view of instrumental rationality. On the standard view, what counts as a coherent set of preferences explicates a fixed standard of ideal rationality. However, by allowing the scope of this demand to be an open, vague, or context-sensitive parameter, we replace this view of rationality with the view that the standards of rationality have some variability, depending upon how we fix this parameter. Second, even if we abandon the inter-contextual demand for coherence, we can raise a similar question with respect to a single decision problem. For any given decision problem, which possibilities ought to be considered? How finely should we describe the states and consequences? So if we are to understand what the strict demand for Deliberative Coherence entails, we must first understand its scope within a single decision problem before we consider more complicated questions about how to extend its scope across decision problems.

The natural first step in making the demand for coherence realistic is to recognize that all-things-considered rationality requires DMs to consider all possibilities even if they are irrelevant. However, when I evaluate whether or not I prefer chicken or fish for dinner, why should the possibility of owning a donkey with a blue hat be relevant? Though all-things-considered rationality demand that I consider everything, this is clearly unnecessary. When we deliberate, we need not consider every possibility, just the relevant ones. So the hope is that we can tailor the demand for Deliberative Coherence to just those possibilities that are relevant for a DM's decision problem. We may do so by embracing

Adequate Coherence: For each decision problem, preferences should incorporate the DM's beliefs and desires about all the relevant states and consequences.

This raises the obvious question: how do we determine what counts as relevant? Why are questions about owning donkeys irrelevant for my choice between chicken or fish? Why are questions about the restaurant's cleanliness relevant? The natural response is that if my choice between meals is a normal one, my preferences should not, by my own lights, be affected by considering questions about owning donkeys. In contrast, my reasonable preferences – those that are justified by my beliefs and desires – could be affected by considering questions about cleanliness. So relevant possibilities are just those possibilities the consideration of which could affect one's reasonable preferences. Adequate Coherence demands that we consider everything that is relevant by our own lights. The result is an

All-Things-Relevant Rationality: For each decision problem, a DM is rational if and only if she has a coherent set of preferences that are determined by the DM's beliefs and desires about all the respectively relevant states and consequences.

A DM satisfies Adequate Coherence so long as she accounts for every possibility whose consideration could reasonably have an effect on the beliefs and desires that she uses to evaluate the choices in a decision problem. Therefore, since rational preferences account for everything that is relevant, the beliefs and desires that determine these preferences should be unaffected were the DM to account for any unconsidered possibility.

The first thing to note is that all-things-relevant rationality abandons the standard account. In the attempt to make Deliberative Coherence applicable, we have abandoned EXTENDED. Adequate Coherence makes no claims about the demands of instrumental rationality outside of our deliberative judgments. It simply proposes that given the possibilities that count as relevant, the corresponding deliberative beliefs and desires must form a coherent set of preferences. But if deliberative judgments are not direct expressions of background attitudes and if there is no applicable demand to have coherent beliefs and desires across different decision problems, then it is not clear how to extend the scope of Deliberative Coherence beyond a particular deliberative context.

More importantly, the demands of all-things-relevant rationality are just as problematic as those of all-things-considered rationality. Of course, in some circumstances one can account for everything that

or are we exhibiting other virtues? One way to make sense of this demand would be to claim that we were getting closer to being all-things-considered rational. However, as I argued in the previous section, the notion of all-things-considered rationality does not apply to human agents.

is relevant. For example, few possibilities seem relevant when choosing whether to open a door with your left or right hand. Unfortunately, such decision problems are special cases. In contrast, think of the possibilities that are relevant for deciding one's next move in a game of chess. One may have to consider many of ways the game could play out, making the decision problem as problematically complex as the all-things-considered decision problem. Or consider the choice of how to spend a weekend. At first glance, I may think that this is a manageable problem. Perhaps I consider the consequences that my choices have on my immediate family. However, this is clearly too narrow of a view. My choices might have an effect on my work and well-being just as it might have an effect on anyone who I would or would not come into contact with. These effects might then have their own repercussions and so on and so forth. To account for everything of relevance, I may have to consider many of ways the world could play out. As Quine argued, when taken in combination with other beliefs, almost any observation could be relevant for the (dis)confirmation of an empirical hypothesis.[20] Similarly, when taken in combination with certain states, almost any consequence could be relevant for the evaluation of an act. Since all-things-relevant rationality requires DMs to grasp and evaluate overly complex propositions, its demands also fails to satisfy the 'ought' implies 'can' maxim.

The problem we face is aptly summarized by Bengt Hansson,

> A practicable theory must make it possible for a gambler to decide how to play a game without deciding his whole future or even all his future games at the same time and likewise for a businessman to decide how much he shall bid for a job without having a complete strategy for his firm. The complete description of the future must be divisible into 'the current problem' plus 'the rest', where the two parts can be decided separately.[21]

Thus, all-things-considered and all-things-relevant rationality run into similar problems of applicability. Not only is there a problem in making the theory applicable when we consider how to extend the scope of Deliberative Coherence across many decision problems, but the problem also remains when we consider the scope of Deliberative Coherence within a single decision problem.[22]

2.3 Local Coherence and All-Else-Neglected Rationality

We are left with two general problems for understanding the normative significance of decision theory and Deliberative Coherence. These problems have arisen from the problems that we have raised for EXTENDED and COMPLETE. First, the problems for EXTENDED raise the general problem of demarcating which possibilities ought to be considered for a single decision problem. And so it is difficult to see for each decision problem which beliefs and desires are governed by Deliberative Coherence. As Hansson noted, we need a way to demarcate what is relevant for the current problem while setting aside the rest. Second, the preference reversal phenomenon undermined COMPLETE. A more comprehensive account of instrumental rationality should thereby govern how DMs construct rational preferences.

I believe that we can address both of these challenges by rethinking the significance of Deliberative Coherence in two ways. First, we must localize the scope of Deliberative Coherence, explicitly restricting its purview to single decision problems. Next, rational preferences and the scope of Deliberative

[20](Quine 1951)

[21](Hansson 1975: 179) Savage famously noted that "it is even utterly beyond our power to plan a picnic or play a game of chess in accordance with [the principles of rational choice] even when the [set] of states and the set of available acts to be envisaged are artificially reduced to the narrowest reasonable limits." (Savage 1972: 16) He attempted to deal with this problem by appealing to small worlds, and the following discussion can be seen as an attempt to explicate what small worlds must be. There are other attempts to address this problem of application (c.f. Gilboa and Schmeidler 2001). Most notably, (Binmore 2008) attempts to offer a minimal extension of Bayesian decision theory to cases in which the DM cannot account for everything of relevance. Nevertheless, Binmore agrees with my conclusion that neither all-things-considered rationality nor all-things-relevant rationality are viable accounts of human rationality. His positive proposal can be seen as an alternative to the one I propose below. However, given the scope of this paper and its focus on incorporating the lessons from the empirical research on preference reversals, I will not discuss his proposal.

[22]In order for a decision theory to be applicable to humans, it need not provide a decision procedure. Instead, it must simply govern deliberative activities that we are capable of engaging in.

Coherence must be context-sensitive. That is, what counts as rational and which beliefs and desires ought to be coherent will depend on context-sensitive features of a decision problem. My proposal is that the features that matter are the parameters that the DM sets for her decision problem. When a DM sets these *decision parameters*, she accomplishes two tasks. First, she demarcates the states and consequences that count as relevant for that decision problem, and Deliberative Coherence only governs the corresponding beliefs and desires. Second, in demarcating what counts as relevant, she also places some constraints on the preferences she can rationally have for that decision problem (i.e., restrict the set of rationally permissible expected utility functions). So decision parameters play a central role in the rational construction of preference.[23] Taken together, the proposal is that the Deliberative Coherence of DM's preferences is limited to a deliberative context and relative to the way she sets the decision parameters. In the next section, I will fill in the details of this proposal.

Bracketing, for the moment, questions about how we should construct the relevant deliberative beliefs and desires, this local, context-sensitive view of Deliberative Coherence endorses

Local Coherence: For each decision problem, rational preferences need only to incorporate the DM's beliefs and desires about only those states and consequences that are demarcated as relevant.

What differentiates Adequate Coherence from Local Coherence is that each appeals to different subjective notions of relevance. Adequate Coherence appeals to an unrestricted notion of relevance. Everything that the DM cares about is relevant. Local Coherence appeals to a restricted notion of relevance. The DM takes some cares and concerns as relevant, neglecting all others. And only those possibilities that are relevant to these limited cares and concerns are accounted for. For example, when deciding where to go eat, Steve may only concern himself with having a cheap meal. Given this concern, Steve can assess his choices in a simple way: the cheaper, the better. Alternatively, Steve may only be concerned to have a quick meal. Given this concern, the faster, the better. In one situation, only the cost is considered while in the other, only the duration of the meal is considered. Adequate Coherence would require the DM to consider both concerns as well as all others. What results from Local Coherence is

All-Else-Neglected Rationality: For each decision problem, a DM is rational* if and only if she has a coherent set of preferences that are determined by her beliefs and desires about respectively those states and consequences that have been demarcated as relevant – everything else is neglected.[24]

Of course, this notion of all-else-neglected rationality does not, by itself, offer an adequate account of instrumental rationality. For this reason, let *rational** denote a minimal type of rationality. In §3.2, I will use this minimal type of rationality to develop a richer and more complete account of instrumental rationality. However, before I do so, I must show how all-else-neglected rationality addresses the two challenges raised by preference reversals by filling in its details. I will sacrifice comprehensiveness for detail by focusing my discussion on the way a DM may demarcate a set of relevant consequences and construct the corresponding deliberative desires. Though the demarcation of relevant states and the construction of deliberative beliefs will depend upon different types of judgments, such as judgments of evidential relevance (e.g., symmetry judgments), there will be structural similarities. For a constructive view of probability judgment that can be used to fill in the account, see (Shafer and Tversky 1985). By elaborating on the details of the constructive process, I hope to highlight the close relationship between the demarcation of relevant consequences and the construction of deliberative desires.

At this point, it is worth making a quick sidenote for those who may be thinking that the problem of demarcating the relevant consequences is a familiar one that has been raised and addressed by a number of writers. For example, John Broome has argued that decision theory has normative content

[23] DMs must also identify which acts are available to perform. This is an important parameter of a decision problem but I shall ignore it for our discussion.

[24] Two interpretations of this view present themselves. On the strong interpretation, for the multitude of decision problems the DM could face, she ought to possess rational preferences for each. On the weak interpretation, the DM only ought to possess rational preferences for the decision problem of current concern. As my discussion in §3.2 will show, I think there are various ways of assessing a DM's rationality. Therefore, being instrumentally rational cannot be reduced to one or the other interpretation.

only if there are principles governing the demarcation of relevant consequences.[25] Broome endorses rational requirements of indifference that are grounded in an external criteria of goodness.[26] While the problems that I have raised and plan on tackling are closely related, there are a few important differences. First, I am interested in both the question of how to demarcate consequences and how to construct deliberative desires. Furthermore, I think that these questions are intimately related. So even if Broome had an adequate answer to the former question, he is simply not concerned with the latter. Second, Broome's claim that decision theory has normative content only if there is some way of demarcating consequences is stronger than one that I am committed to.[27] I have argued for the weaker claim that theories, even if they possess normative content, do not apply to bounded agents unless we restrict the consequences that matter.

3 Goal-Relative Rationality

To fill in the details of all-else-neglected rationality, I will answer the following questions.

1. How may the DM restrict the consequences that she considers as relevant to her decision problem?
2. How are a DM's deliberative desires related to her background attitudes and the deliberative context?
3. How should a DM construct her deliberative desires?

My answers come from a reflection on the role that goals play in decision making. When deciding between taking the subway or a taxi during rush hour traffic, I may deliberate on the assumption that I want to make it on time to a meeting. On a trip to the casino, a gambler may decide that she is going to win at least three hundred dollars or go home empty handed. In each case, the adopted goal identifies a standard for evaluating the consequences of one's choices. The consequences of taking the subway or train will be evaluated relative to how well it achieves the goal of making the meeting on time. The consequences of the gambler's strategies will be evaluated relative to whether she wins at least three-hundred dollars. By focusing on the way goals specify a standard for evaluating the consequences of our choices, we will be able to show that through the adoption of a goal for one's deliberation, the DM may restrict the set of relevant consequences and identify a standard that should be used to evaluate the desirability of these consequences.

3.1 Goals: Values to Preferences

In order for a goal to specify a standard of evaluation, one must identify what is means to satisfy a goal. For example, the goal of arriving on time for a meeting may be a goal that one satisfies if and only if one arrives on time. Moreover, no extra details about the consequences of one's actions matter relative to this particular goal. It wouldn't matter whether one is sweaty or loses a shoe in the process. If the

[25] "There *must* be such [principles], if rationality is to constrain practical preferences in any way." (Broome 1995: 105) (Anand 1995), (Schick 1991), (Schick 1997) also discuss related issues. (Bermudez 2009) has a long discussion of individuating consequences but is interested in a different inquiry about whether we can make compatible the internal and external perspectives of decision theory.

[26] Broome thinks that there are "rational requirements of indifference, which limit the fineness of individuation that is allowed."(Broome 1995: 104) And "it is rational to have a preference between two alternatives only if they differ in some good or bad respect." (Broome 1995: 106) I will appeal to the DM's values to demarcate consequences but no external criteria of goodness will constrain which values may be relevant.

[27] Broome's conclusion depends upon the assumption that there is always a way of ensuring the coherence of preferences by simply individuating the relevant acts in a more-fine grained way. However, this assumption is questionable. Suppose that we have acts A', B', and C' that are respectively assumed to be refinements of acts A, B, and C. As (Savage 1972) argued, coherence demands that the DM's preferences over the first set of acts mirror the DM's preferences over the second set. Therefore, if the DM has a set of transitive preferences over the first, but a set of intransitive preferences over the second, we cannot restore coherence by finely individuating the latter set of acts – so long as we assume that the latter are indeed refinements of the former.

goal is merely to make it on time, there is only one way to satisfy that goal. To put it another way, relative to this goal, there is only one potential property that is relevant for describing and assessing the consequences of the DM's actions, the property of being on time.[28] Of course, our goals and what it means to satisfy the goal can be more fine-grained. Given the goal of running a marathon quickly, there are better or worse ways of accomplishing this goal. So the DM should consider a more fine-grained set of mutually exclusive and exhaustive properties that describe how quickly she runs the race, and the relevant consequences are those that are individuated by these properties.

So when we adopt a goal, we must articulate what it means to satisfy that goal. In articulating what it means to satisfy that goal, we identify a set of relevant properties. And the set of relevant consequences are any of those consequences that can be picked out by the relevant properties. Therefore, *goals determine the relevant consequences.*

Goals also determine which of the relevant consequences ought to be better or worse. Relative to the goal of running the marathon quickly, the faster time ought to be better. Goals play this role in part because they demarcate a set of relevant values.[29] For our purposes, *values* should be understood as desires or values for properties and can be represented by ceteris paribus preferences.[30] These are preferences over consequences or states of affairs rather than acts, and they can be used to identify the relative desirability of consequences.[31] In the simplest cases, there is only one relevant property. So the value for being punctual to a meeting is represented by a ceteris paribus preference for consequences that instantiate the property of the DM being on time over consequences that do not instantiate this property. The ceteris paribus clause means that this is a preference the DM has only if every other detail about these consequences is identical.

While these simple values offer a very coarse-grained evaluation of possible states of affairs, our values can provide more fine-grained assessments as well. Relative to the value of running a marathon quickly, what is relevant is a richer set of properties that specify how quickly the race is run. This value can be represented by a set of ceteris paribus preferences over the consequences individuated by these properties. Moreover, values provide a complete ranking over the relevant states of affairs. My value for running the marathon quickly entails that for any two distinct consequences that describe how fast I run, I have a ceteris paribus preference for the consequence where I run faster.

Multiple values may also be associated with a goal. In adopting the goal of implementing the best city-wide health initiative, I may take the best initiative as that which best improves the city's health and is the most cost effective and politically viable. As a result, the properties that describe a city's health as well as the cost-effectiveness and political viability of the potential initiatives are relevant. Associated with this goal are the values for each respective set of properties.

When more than one value is relevant, these values alone do not determine the relative desirability of all the relevant consequences. For example, the value for wealth and leisure are not enough to determine whether I should prefer a state of affairs where I am wealthy but lack leisure to another state of affairs where I enjoy leisure but lack wealth. Since these values only provide ceteris paribus preferences and the two states of affairs are not equal with respect to all the relevant features, these values leave us with an indeterminate preference. In decision making contexts, what allows us to assess the trade-offs that we are willing to make between values (i.e., the relative importance of each value) is the adopted goal. So *goals can be used to determine a preference when the relevant values conflict.*

For instance, I may adopt the goal to have my weekends free to spend with family and friends. And so long as this is satisfied, I will do whatever I must to build as much wealth as possible. Given this goal, I should prefer the consequence where I have less wealth but have my weekends free over one where I possess greater wealth but spend my weekends working. Of course, I may adopt another goal

[28] Following Savage, consequences can be understood as states of the agent, and they pick out what the DM's experiences as a result of her actions. So these properties are instantiated by the DM.

[29] The proposed view of goals, values, and properties is influenced by the discussion in chapter 2 of (Raiffa and Keeney 1976).

[30] For a discussion of desires as values for properties as opposed to states of affairs, see (Pettit 1994)

[31] Using *constant acts* – acts that result in the same consequence regardless of what the world is like – we can talk about preferences for consequences by appealing to our preferences between acts.

that takes the same values as relevant yet articulates different trade-offs. Suppose I adopt the goal of creating as much wealth as possible so long as I am able to enjoy at least one meal a week with my friends and family. The same values may be in play, but with respect to this goal, I should weigh the value for wealth more heavily. Therefore, goals determine a set of relevant values, but a set of values does not entail a goal.

In summary, goals are partial functions from a set of values to a set of deliberative desires.[32] Figure 1 offers a pictorial representation of these partial functions. We can use this diagram to summarize the goal-relative construction of deliberative desires and to answer the three questions that I raised above.

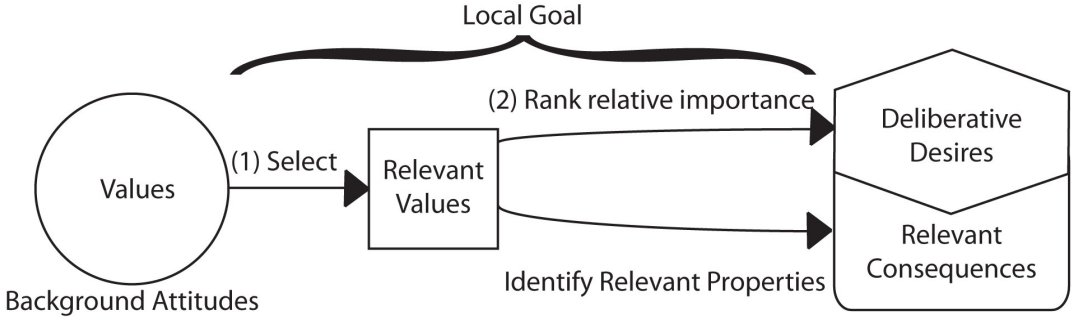

Figure 1: Construction of Deliberative Desires

Values are fairly stable features of a subject's state of mind. Most of us generally value wealth, leisure, and health. And we possess these values regardless of the situation that we are in. Given their stability, it is plausible to view values as an important part of a DM's background attitudes. Next, there is an indirect relationship between a DM's values and her deliberative desires. In any given decision problem, not every value may be relevant. There may be instances where the DM ignores the effect of her choices on her overall wealth and other instances where that value is of primary importance. Goals, as I have described them, are adopted for a decision problem and play the intermediary role between values and deliberative desires. They interact with our values in two ways to produce deliberative desires.

1. Goals select the values that count as relevant for a decision problem, which in turn is sufficient to determine the set of relevant consequences.

2. Goals determine the relative importance of these values, which in turn is sufficient to determine the DM's deliberative desires for a decision problem.

Consider the following example. I plan to run a marathon. However, I have not properly trained so I am worried about my physical health. Before the race, I adopt the following goal: I will do everything in my power to finish so long as I do not endure any lasting physical damage. Figure 2 summarizes the constructive use of this goal.

From all the values that I possess, my goal selects two relevant values, the value for success and for health. For simplicity, let's suppose that I consider a fairly coarse-grained set of properties. Given my value for success, I consider two properties: finishing the marathon or not finishing. Given my value for health, I consider the properties of being healthy, experiencing temporary pain, and having permanent physical damage. As a result, there are six possible consequences that can be picked out, and relative to the values for success and health, a is best, e is second worst, and f is worst. Though

[32]While the higher-order desires and meta-preferences discussed respectively by (Frankfurt 1971) and (Sen 1977) can function as goals, they are not the only attitudes that can play this role. In fact, any attitude, judgment, or combination of attitudes and judgments that can select a set of relevant values and determine one's preferences over the relevant consequences can function as a goal.

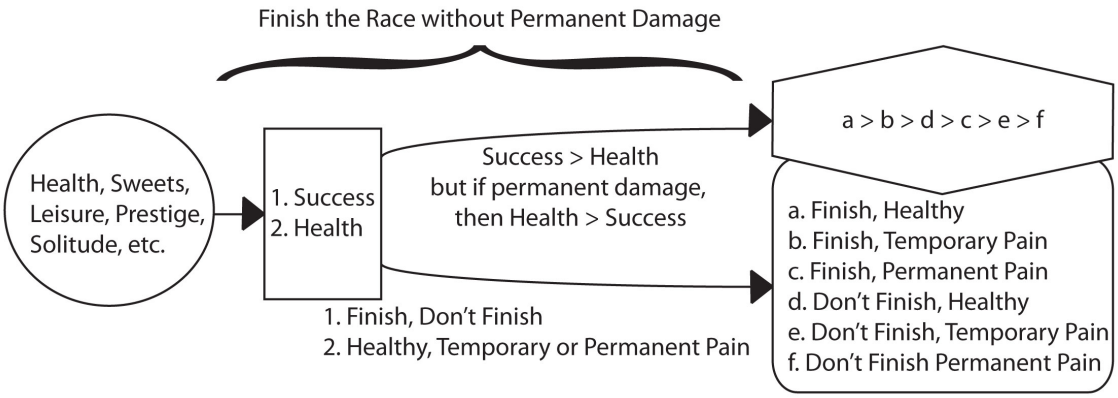

Figure 2: Constructing Deliberative Desires Example

I know that the remaining consequences lie somewhere between a and e, the selected values alone do not determine their relative desirability. For what is more desirable, finishing and enduring permanent physical damage or not finishing and being in good health? Here, the adopted goal specifies the relative importance of these values and can thereby be used to determine my preferences. Since success is more important than health only when I do not endure permanent damage, I should rank b, d, and c in a descending order of desirability.

By representing goals as partial functions from values to deliberative desires, goals are able to place constraints on the desirability of consequences. If one's goal is to finish the marathon no matter what, then it would be irrational to find it more desirable to prevent the mildest physical pain over the successful completion of the marathon. This assessment of desirability is irrational in the sense that it is not permissible or coherent given the adopted goal. So goals play a constructive role by restricting the set of rationally permissible deliberative desires. To state this another way, on the orthodox Bayesian account, so long as one's desires are representable by some utility function, they count as rational. But on the goal-relative account, goals place additional constraints by restricting the set of rationally permissible utility functions that may represent our desires.[33] And one's deliberative desires are rational only if they are rationally permissible given the adopted goal. Since we haven't discussed the rational construction of deliberative beliefs, the account is incomplete. So for the sake of simplicity, let's assume that there is a way to restrict the set of relevant states and that any probability measure over these states constitutes a set of rational deliberative beliefs. Therefore, goals will only place rational restrictions on our deliberative preferences by restricting the set of rational deliberative desires. The result is a particular version of all-else-neglected rationality.

Goal-Relative Rationality: For each decision problem, a DM is rational* if and only if she has a coherent set of preferences that is rationally permissible given the adopted goal.

3.2 Higher-Order Rationality: Myopia and Preference Reversals

On the standard decision-theoretic account, the entirety of a DM's background beliefs and desires ought to form a coherent set of preferences. So long as there is no change in what she believes or desires, these preferences should then determine a suitable ranking of acts for every decision problem. In contrast, all-else-neglected rationality proposes that the preferences that count as rational* may change from one deliberative context to the next depending upon the parameters the DM sets for her decision problem. The goal-relative account fills in some of these details, proposing that the DM's local goals are part of what sets the parameters of her decision problems. However, were that the end of the story, we

[33]For simplicity, I've assumed that a goal entails a unique rationally permissible set of complete preferences over the consequences. This assumption can be weakened.

would be left with an overly narrow view of instrumental rationality whose demands are clearly too weak. Fortunately, the minimal type of rationality explicated by goal-relative rationality can be used to develop a much fuller account of instrumental rationality, one that is much richer than the standard view.

In addition to answering the challenges raised by preference reversals, the goal-relative account allows us to distinguish between first-order and higher-orders of rationality. A DM is first-order rational if and only if she has a coherent, appropriately constructed set of preferences for her decision problem. Suppose Steve is deciding where to eat and has a coherent set of preferences reflecting a simple concern for his health. Now, Steve can step back from this choice between dining options and consider the adoption of a goal as a decision problem in and of itself. Should he aim to be healthy, be efficient, or both? In doing so, he is taking a second-order approach to his decision problem. Steve is second-order rational if and only if he has a coherent, appropriately constructed set of preferences with respect to this *higher-order decision problem*. These are decision problems about how to fix the parameters of another decision problem. We could even consider a third-order decision problem. Consider a number of first-order decision problems about where to go eat for lunch. In turn, each of these can be associated with a second-order decision problem about what goal to adopt when deciding where to go eat. The third-order decision problem might then be concerned with what the DM should be aiming for when deciding which goals to adopt for each first-order problem. We might characterize these as deliberations about choosing between strategies for evaluating lunch options.

So goal-relative rationality supplies the basic building block in the construction of a full account of instrumental rationality. We begin with the minimal type of rationality that is localized to particular decision problems and sensitive to the way we set decision parameters. This allows us to accommodate the fact that we can deliberate about how to set these decision parameters. As a result, goal-relative rationality allows us to distinguish various orders of instrumental rationality. The resulting recursive and hierarchical structure to our assessments of rationality can be used to address objections to and highlight features of the proposed account. It's important to note is that by adding this structure, it will no longer make sense to simply say that a DM is or is not rational. For a DM may be first-, second-, or third-order rational. In addition, DM's that are third-order (ir)rational may be (ir)rational with respect to fairly small or very large classes of decision problem. So there are a variety of ways in which we could assess the extent to which a DM is (ir)rational.

Let us now turn to consider a variety of objections to the proffered view of instrumental rationality. The obvious objection to goal-relative rationality comes from the irrationality of myopic DMs. The minimal notion of rationality allow DMs to view their decision problems in very restricted ways even when it would be of almost no additional cost to deliberate in a more comprehensive way. So the following example appears to undermine the view. Suppose Steve is asked to evaluate two bets. Though the expected monetary gain is identical for both bets, for one of the bets he is guaranteed to be flogged. It seems obvious that Steve should prefer the bet for which he will not be flogged. However, according to the goal-relative account, Steve may be perfectly rational to be indifferent between the two bets. For we can imagine two different goals that Steve might adopt. And on the view that I have proposed, Steve may adopt any goal for his decision problem so long as it is coherent with his preferences. So he may adopt the goal of making the most money possible or the goal of making the most money possible while minimizing physical pain. If Steve adopts the former goal, his indifference is rational.

It is natural to think that myopic DMs exhibit some type of irrationality, and at first glance it may appear that goal-relative rationality precludes us from saying so. However, while Steve is first-order rational, he is clearly second-order irrational. If he were to deliberate about which goal to adopt, he should prefer to deliberate for the sake of the broader more inclusive goal. Given the options that are available, there is no way the world could turn out in which he would do worse by his own lights if he adopted the broader goal. Of course, if the choices were different, the broader goal may not obviously be better. What if one bet offered the opportunity for untold wealth while the other bet offered the opportunity for a sustained state of euphoria. Would it be better to aim for riches, physical bliss, or both?

The case of myopia highlights the need to differentiate different types of rationality. It would be strange to think that myopic DMs are irrational in just the same way as DMs with explicitly incoherent preferences. If I prefer visiting London over Dubai and Dubai over Singapore, but also prefer visiting Singapore over London, I exhibit a first-order irrationality. My preferences are incoherent, and a symptom of this incoherence is that it is difficult to make sense of them. It is hard to identify what values these preferences are revealing. Call DMs who are first-order irrational, *unintelligible*. In contrast, the preferences of myopic DMs, like Steve, make sense. For we can certainly imagine evaluating actions only in terms of their expected monetary gain while ignoring the possibility of physical pain. Given this distinction, we can properly identify the sense in which myopic agents are irrational. Myopic agents are those who lack a certain reflectiveness about their lives. It's not that they fail to properly exercise a basic capacity for means-ends reasoning. Rather, they fail to appropriately exercise this capacity with respect to higher-order questions about what goals to adopt. The goal-relative account allows us to view these deliberative activities as both distinct and similar. Though each activity is governed by the norms of instrumental rationality, we can distinguish these deliberations as being at different levels. And deliberating rationally in one decision problem does not mean that you have deliberated rationally in the other.

This brings us to the general objection that the proposed account is too lenient. Goal-relative rationality is ensured to be a practicable theory simply because it allows the DM complete freedom in separating what is relevant to 'the current problem' from what is part of 'the rest'. But why should we think that the DM is rational to limit her concerns in any way she likes. If a DM is only concerned for her health when choosing where to eat, surely she should still recognize that other concerns may be of relevance. My response is two-fold. First, if the theory is to be applicable to human DMs, then the DM must be allowed to restrict what counts as relevant. As I argued, human DMs cannot take everything into account. They cannot even take everything of relevance into account. Next, by appealing to the hierarchical and recursive nature of deliberation, the account can capture the intuition that even though DMs must limit what is of current concern, they nevertheless ought to recognize that unconsidered possibilities could be relevant. That is, they ought to recognize the relevance of these unconsidered possibilities from the perspective of a higher-order deliberation.

To show how, let's consider an example. Suppose Steve adopts the simple goal of being comfortable when choosing what to wear for the day. Unexpectedly, the woman he has been pining after turns up at his work. The moment he sees her, Steve thinks, "I was so short-sighted! I should have dressed differently." Given Steve's goal, he was first-order rational to dress as he did. So how can we capture the intuition that Steve's concern for romance, though neglected, should have been relevant when deciding what to wear? We can do so by considering the second-order choice problem in which Steve chooses a goal to adopt for deciding what to wear. What Steve regrets is making comfort his sole concern. Additional values are relevant in the sense that they could have been accounted for by adopting a different goal. And if Steve's regret is reasonable, then there must be a sense in which he was irrational to have dressed the way he did. On my account, Steve may regret his narrowly rational choice if and only if from the perspective of the higher-order deliberation, a different goal would have been best, and on the basis of this alternative goal, a different choice of wardrobe would have counted as best.

Another objection worries about the hierarchy of deliberations. In principle, DMs could always adopt a higher-order perspective of their decision problems but if there is no end to this hierarchy, then a fully rational agent would be left with the impossible task of endlessly ascending this hierarchy.[34] The first and most important point to make in reply is that it is an open question whether instrumental rationality demands that we "Be as rational as possible" or "Be rational to the nth-order." I, myself, am skeptical that there are such additional demands. However, since a full discussion of this is beyond the scope of the paper, let me simply ease the worry by offering a potential resolution. Suppose that we always ought to engage in higher-order deliberations. In practice, DMs can adopt self-justifying goals that stop the regress. Consider a decision problem for which Steve adopts the goal of being a good son. Further, let us suppose that Steve is committed to adopting the same goal in every higher-order

[34]For further discussion of this problem, see (Smith 1991), (Velleman 1997), (Lin 2014).

decision problem. As a result, for each decision problem about what goal to adopt, this would entail that choosing the goal of being a good son is the best option.[35] So this goal would be used to justify its own adoption in every higher-order choice problem. Such goals can be characterized as ones that are taken to be intrinsically valuable or ends-in-themselves for the sake of the lowest level decision problem for which they are adopted.

In order to highlight the explanatory and normative value of differentiating the various levels of rational evaluation, let us return to preference reversals. On the goal-relative account, we can offer a story on which these reversals are first-order rational. In CHOICE, a DM could have deliberated relative to the goal of saving money for the future, while in MATCH, adopted the goal of continuing her world travels.[36] Moreover, this shift in goals between related decision problems may be perfectly rational from a second-order perspective. For there is nothing irrational about deciding to adopt one goal in one decision problem while deciding to adopt another inconsistent goal in another. However, such shifts do appear irrational relative to a third-order decision problem. And in this case, the third-order decision problem concerns what the DM should be aiming for when deciding which goals to adopt for both CHOICE and MATCH. This is a choice between strategies for evaluating job offers. And if the DM reverses preferences between the same choices when presented with those choices in different ways, as in CHOICE and MATCH, she clearly has not adopted a coherent strategy.[37] So we might call DMs that are third-order irrational, *flip-floppers*.

4 Conclusion

I have proposed that the goal-relative notion of instrumental rationality solves two important problems and enriches our view of instrumental rationality. First, by localizing the demands of Deliberative Coherence to particular decision problems described in restricted ways, we ensure that the theory is applicable to human agents. In fact, I have argued for the stronger conclusion that we must localize the demands of Deliberative Coherence or else they will fail to be applicable to human agents. Second, by relativizing minimally rational preferences to goals, we accommodate the construction of preference. Finally, by appealing to a minimal goal-relative rationality and positing a recursive, hierarchical structure to the way we can apply this minimal notion, we are left with a richer view of instrumental rationality.

To conclude, let us contrast the proffered view of instrumental rationality with the standard decision-theoretic account, which was characterized by the commitment to COMPLETE and EXTENDED.[38]

[35]The adoption of a goal is not always the best means of achieving it. For example, Steve may be his worst enemy, perpetually engaging in self-sabotage. Knowing this about himself, he may know that by aiming to do A, he will most likely not achieve A. Though more discussion is needed, the simplest reply is that self-justifying goals are typically but not always available to rational agents. And there may be something pathological about agents who cannot adopt self-justifying goals.

[36]Of course, it may be difficult to identify what these different goals might be. An interesting case study comes from a published interview (see pp. 65-68, Lichtenstein and Slovic 2006) where the subject recognizes some inconsistency between her preferences but refuses to change them. Given her own reports, it appears that she justifies this inconsistency by taking herself to be involved in different tasks with different aims in the two decision problems.

[37]The appeal to third-order coherence can be used to explain the one situation in which preference reversals clearly disappear. These are cases in which subjects are forced into market arbitrage (Chu and Chu 1990). This simply means that they are put through a money pump. In these cases, if the subjects do not want to lose money, they are forced to engage in the third-order decision problem and must adopt a coherent strategy for evaluating bets across various decisions problems.

[38]While it would go beyond the scope of this paper to offer a detailed discussion, I'd like to make a small remark about the view of preference found in (Hausman 2012). Hausman argues that the preferences appealed to in formal economic models are best understood as stable, total comparative evaluations. These are all-things-considered judgments. He notes, however, that the preferences we reveal in our actual choice behavior are typically context-sensitive. Hausman proposes that we can respond in two ways. First, we can separate choice from preference and develop theories of choice that do not appeal to total comparative evaluations. Second, we can retain the connection between choice and preference if we adopt a different view of preference. It is typically thought that by adopting the second way, our theories of actual choice are divorced from theories of rational choice. While I have attempted to take the latter route, I also hope that the proffered view of instrumental rationality will be able to retain some connection between theories of rational choice and theories of actual choice.

While I have rejected the claim that instrumental rationality is just a matter of coherent beliefs and desires, I have retained the more general Humean thesis that instrumental rationality is just a matter of coherence. The impetus for this revision is that in order to account for the rational construction of deliberative beliefs and desires, we should incorporate goal judgments into the set of judgments that must be coherent. I have also rejected the simple thesis that the demand for Deliberative Coherence can be directly extended to govern our background attitudes. However, I do think that instrumental rationality can govern our background attitudes. Background attitudes are simply those attitudes that remain somewhat stable across decision problems. So if there are demands of coherence across decision problems, then we can impose demands for coherence to background attitudes. Previously, I had rejected any fixed, absolute demand to be coherent across decision problems. However, I think that we can and should replace the absolute demand with a set of more nuanced ones that are associated with different attributions of rationality. On the standard account, we can only make simple attributions of rationality.[39] However, it seems clear that there are varying degrees to which DMs are or are not rational, and my account is able to differentiate narrow assessments of rationality from more global assessments of rationality, which are relevant for evaluating background attitudes. As we consider more global assessments of rationality, we evaluate deliberative judgments that are stable across more and more decision problems. The unintelligible DM fails to be rational in the minimal sense. The myopic DM, though intelligible, fails to satisfy a type of reflective rationality, one that we expect reasonable people to satisfy. Flip-floppers are more difficult to assess. Though they are irrational in some sense, is it a reasonable demand that people be coherent across various choice problems?[40] I suspect that the answer might depend upon many factors like the relationship between the various choice problems. Whatever we conclude, the very fact that these distinctions and questions can be raised highlight the upshot of adopting the view of instrumental rationality that is grounded in goal-relative rationality. For being rational is less like being a member of a club and more like being competent. And in the latter case, we naturally ask, how competent and for what purpose? We should be able to ask similar questions about being rational.

References

Allais, Maurice (1979). *The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School (1952)*. Springer.

Anand, P., P. Pattanaik, and C. Puppe (2009). *The Handbook of Rational and Social Choice*. OUP Oxford.

Anand, Paul (1995). *Foundations of rational choice under risk*. OUP.

Bermudez, Jose Luiz (2009). *Decision Theory and Rationality*. Oxford: Oxford University Press.

Binmore, Ken (2008). *Rational Decisions*. Princeton University Press.

Broome, John (1995). *Weighing Goods*. Oxford: Wiley-Blackwell.

— (2013). *Rationality through reasoning*. John Wiley & Sons.

Chesterton, G.K. (2012). *The Defendant*. Dover Publications.

Chu, Y.P. and R.L. Chu (1990). "The subsidence of preference reversals in simplified and marketlike experimental settings: A note." *The American Economic Review* 80.4, pp. 902–911.

Cubitt, R.P., A. Munro, and C. Starmer (2004). "Testing explanations of preference reversal." *Economic Journal*, pp. 709–726.

Fantl, J. and M. McGrath (2002). "Evidence, pragmatics, and justification." *The Philosophical Review* 111.1, pp. 67–94.

[39] For example, Bermudez writes that "we need a way of individuating outcomes that will determine in an absolute sense whether or not the action is rational." (Bermudez 2009: 103) I am suggesting that we abandon the view that decision theory provides an absolute standard of rationality.

[40] G.K. Chesterton offers a poetic argument against flip-flopping, characterizing flip-floppers as people of decadence whose vice is that they desire to be many different people. What is interesting about Chesterton's discussion is that even though he argues that flip-floppers fail to experience the joys that can only come from unwavering commitment, he nevertheless admits that the view of the decadent "is a perfectly possible, rational and manly position." p. 12, (Chesterton 2012)

Fischer, G.W. et al. (1999). "Goal-based construction of preferences: Task goals and the prominence effect." *Management Science*, pp. 1057–1075.
Frankfurt, Harry G. (1971). "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68.1, pp. 5–20.
Gibbard, Allan (2009). *Thinking how to live*. Harvard University Press.
Gilboa, I. and D. Schmeidler (2001). *A theory of case-based decisions*. Cambridge Univ Pr.
Hansson, Bengt (1975). "The Appropriateness of the Expected Utility Model." *Erkenntnis*, pp. 175–193.
— (1988). "Risk aversion as a problem of conjoint measurement." In: *Decsion, Probability, and Utility*. Cambridge Univ Press.
Hausman, Daniel M (2012). *Preference, value, choice, and welfare*. Cambridge University Press.
Jeffrey, Richard (1965). *The Logic of Decision*. Chicago: University of Chicago Press.
Joyce, J.M. (1999). *The Foundations of Causal Decision Theory*. Cambridge Univ Press.
Kim, Brian (2012). "The Context-Sensitivity of Rationality and Knowledge." PhD thesis. Columbia University.
Levi, Issac (1999). "Value commitments, value conflict, and the separability of belief and value." *Philosophy of Science*, pp. 509–533.
Lichtenstein, S. and P. Slovic (1968). "Relative importance of probabilities and payoffs in risk taking." *Journal of Experimental Psychology* 78.3p2, p. 1.
— (1971). "Reversals of preference between bids and choices in gambling decisions." *Journal of experimental psychology* 89.1, pp. 46–55.
— (2006). *The Construction of Preference*. Cambridge Univ Press.
Lin, Hanti (2014). "On the regress problem of deciding how to decide." *Synthese*, pp. 1–10.
Pettit, P. (1994). "Decision Theory and Folk Psychology." In: ed. by M. Bacharach and S.L. Hurley. Blackwell. Chap. 4, pp. 147–175.
Quine, WV (1951). "Main trends in recent philosophy: two dogmas of empiricism." *The Philosophical Review* 60.1, pp. 20–43.
Raiffa, H. and R. Keeney (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley.
Savage, L.J. (1972). *Foundations of Statistics*. New York: Dover.
Schick, Frederic (1991). *Understanding Action: an essay on reasons*. Cambridge University Press.
— (1997). *Making choices: A recasting of decision theory*. Cambridge University Press.
Sen, Amartya K (1977). "Rational fools: A critique of the behavioral foundations of economic theory." *Philosophy & Public Affairs*, pp. 317–344.
Shafer, G. and A. Tversky (1985). "Languages and designs for probability judgment." *Cognitive Science* 9.3, pp. 309–339.
Shafer, Glenn (1986). "Savage Revisited." *Statistical Science* 1.4, pp. 463–485.
Simon, H.A. (1955). "A behavioral model of rational choice." *The quarterly journal of economics* 69.1, pp. 99–118. ISSN: 0033-5533.
Smith, Holly (1991). "Deciding How to Decide: Is There a Regress Problem?" In: *Essays in the Foundations of Decision Theory*. Oxford: Blackwell.
Velleman, J. David (1997). "Deciding How to Decide." In: *Ethics and Practical Reason*. Oxford University Press, pp. 29–52.
Wallace, Jay (2008). *Practical Reason. http://plato.stanford.edu/entries/practical-reason*.
Wedgwood, Ralph (2007). *The Nature of Normativity*. Oxford University Press.
— (2011). "Instrumental Rationality." *Oxford Studies in Metaethics* 6, pp. 280–309.