Deliberation and Prediction

YANG LIU AND HUW PRICE
University of Cambridge

Can an agent hold a meaningful credence about an upcoming action, while she deliberates about what to do. Can she believe that it is, say, 70% probable that she will do A, while she chooses whether to do A? Following Spohn and Levi, some writers claim that such 'act credences' are incoherent – Deliberation Crowds Out Prediction (DCOP), as Levi put it. Others claim that the case for DCOP is weak, or even that it is clearly false. We argue these disagreements are to a large extent terminological, or model-relative. After explaining why DCOP does hold in the kind of operationalist model of credence we owe to Ramsey, we note that it is a trivial matter to extend this model so that DCOP no longer holds, in the extended model. We then discuss in detail the model proposed by James Joyce, who presents it explicitly in opposition to Levi and to DCOP. We show that Joyce's disagreement with writers such as Ramsey and Levi rests largely on different choices concerning the use of terms such as 'evidence', prediction' and 'belief'. Once these differences are in view, they reveal a great deal of underlying agreement. In particular, a principle that Joyce calls the Evidential Autonomy Thesis (EAT) is effectively DCOP, in new terminological clothing. We close by proposing that the origin of what is correct in both principles lies in the so-called 'transparency' of the first-person present-tensed view of certain of one's own mental states.

1. Introduction

There is a long-standing disagreement in decision theory about whether an agent can hold a meaningful credence about an upcoming action, while she deliberates about what to do. Can she believe that it is, say, 70% probable that she will do A, while she chooses whether to do A? Following Spohn (1977) and Levi (1989, 1996), many writers have claimed that such 'action-credences' are somehow problematic, or even incoherent – Deliberation Crowds Out Prediction (DCOP), as Levi himself put it. But opinions differ widely about DCOP. Some writers, including Spohn and Levi themselves, take it to be almost a platitude;[1] others (e.g., Ahmed, 2014; Hájek, 2016; Joyce, 2002; Rabinowicz, 2002) think that the case for it is weak, or even that it is clearly false.

1. "Probably anyone will find it absurd to assume that someone has subjective probabilities for things which are under his control and which he can actualize as he pleases," as Spohn (1977, 115) puts it.

As in other such cases in philosophy, it is natural to wonder whether there is a single 'it' at stake. Perhaps the two sides have different principles in mind, and their disagreement is therefore terminological, or partly so. We find this diagnosis, and the irenic resolution of some of these disagreements that it would permit, to be plausible and attractive, respectively. One of our aims in this piece is to clarify the landscape, in support of this peace-making project. In particular, we want to identify several points at which such 'merely verbal' disagreements may be arising.

Given these irenic goal, we must be careful not to speak of *the* DCOP thesis. There may be several theses in the neighbourhood, depending on what one means by terms such as *credence, action, deliberation, choice,* and the like. With this in mind, we proceed as follows. We review one well-known framework for understanding some of these terms, namely, the classical subjective decision theory (SDT) that descends from writers such as Ramsey and Savage. We point out that within this Ramseyian framework, one can find a clear basis for a DCOP-like thesis. This observation is not new, but it is not as well-known as it should be, and we don't know of any previous writers who put its significance into the broader context that we offer here.

The observation shows there is simply no place for action-credences within Ramseyian SDT, *as it stands.* But the qualification is important. We shall see that there are modifications of SDT that do admit action-credences. We focus in particular on a model proposed by James Joyce. Among its advantages for our purposes is the fact that Joyce presents it explicitly in opposition to Levi's defence of DCOP.

Thus Joyce takes himself to be disagreeing with Levi about DCOP (disagreeing rather vigorously, as we shall see). But in our view he misrepresents the nature of the dispute. Comparing Joyce's model to Ramsey's, we shall see that the main difference is that Joyce treats as 'belief-like' some components of the decision process that for Ramsey simply live in a different box altogether – in the intention box, rather than the belief box. Both sides agree that there are such items, and that they can have some of the formal properties of credences – a degree and a propositional content. The disagreement, to the extent that there is one, is about whether they deserve to be called 'beliefs', but this is a terminological matter. (Joyce appears to agree with Ramsey that there are no act credences *in Ramsey's sense* during deliberation.) We propose that Joyce's disagreement with Levi has the same largely terminological character.

We don't deny that there is room for argument about the terminological matter in question (i.e., roughly, whether to treat intentions as a special kind of belief). But our main concern will be to emphasise *similarities,* not *differences,* between Ramsey and Joyce. In effect, we want to show that a rather shallow, terminologically-grounded disagreement about DCOP has obscured a deep point of agreement between the two models, in the form of a principle that Joyce calls the *Evidential Autonomy Thesis* (EAT) – as he formulates it, the thesis that "a rational agent, *while in the midst of her deliberations,* is in a position to legitimately ignore any evidence she might possess about what she is likely to do." (Joyce, 2007, 556–557)

In our view, EAT turns out to be more fundamental than DCOP, while embodying much of what recommends DCOP to its proponents. Thus the arc of our paper bends

towards reconciliation in a deeper sense: not only is the disagreement about DCOP much shallower than usually assumed, but there is a clear prospect of agreement on a more fundamental characteristic of agency.

The remainder of the paper goes like this. In §2 we review the SDT framework, as it derives from Ramsey, with a particular focus on the conception it offers of what credence *is*. In §3 we explain why, within this framework, there is no role for credences concerning a certain class of actions – specifically, credences concerning those actions that SDT takes to be the manifestations of credence in general, at the time at which those actions are playing this revelatory role. (As this formulation hints, the point turns in part on self-reference.)

As we stress, we take this conclusion to be model-relative, and in §4, following a line of thought offered by Joyce (2002), we describe two motivations that have been offered for extending the Ramsey model so as to admit action-credences. One motivation takes the view that action-credences are needed for formal reasons, even though they cannot be manifested directly in the way that SDT takes to be definitional for credences in general. The other claims that action-credences are needed in an adequate model of the deliberative process itself. We note a well-known response to the first consideration, but our main interest at this point will be in showing how little the second conflicts with the spirit of Ramsey and Levi. In other words, our interest will be not in disagreeing with Joyce but in highlighting the respects in which he clearly agrees with Ramsey – here Joyce's EAT plays a central role.

Exploring this point of agreement further, we close with a discussion of why, even if we extend SDT to admit action credences, such credences cannot provide *reasons* for actions. (Joyce himself agrees that this is so.) This turns out to be closely connected to the question as to what grounds EAT itself. We close with the proposal that EAT is a consequence of more general considerations, widely discussed elsewhere – in particular, of the so-called 'transparency' of first-person present-tensed reflection on certain of one's own mental states. Transparency seems to be the key to what Ramsey and Joyce turn out to have in common, once the shallow disagreement about DCOP has been set to one side.

2. Ramsey's Operational Psychology

Modern subjectivists understand credence, or 'subjective probability', in terms of its role in rational decision making. For our purposes, we want to think of this approach as providing a functional *definition* of credence – in effect, credence is treated as a theoretical notion, which is operationally defined, along with subjective utility, in terms of its role in producing certain specified choices. The idea of formalising the notion of credence, or degree of belief, in this way goes back to Frank Ramsey's ground-breaking work 'Truth and Probability' (Ramsey, 1926).[2]

2. Different notions of subjective probability appear earlier in, e.g., Bernoulli (1713), Laplace (1810), **?**, and Borel (1924). But Ramsey is usually credited as the first to provide a systematic account of subjective probability – one of his great contributions being to show that degrees of belief

Ramsey sets out to investigate what he calls "the logic of partial belief," and to treat such a logic as the basis for an understanding of probability. He notes a large obstacle in the path of this project, however:

> It is a common view that belief and other psychological variables are not measurable, and if this is true our inquiry will be vain; and so will the whole theory of probability conceived as a logic of partial belief; for if the phrase 'a belief two-thirds of certainty' is meaningless, a calculus whose sole object is to enjoin such beliefs will be meaningless also. Therefore unless we are prepared to give up the whole thing as a bad job we are bound to hold that beliefs can to some extent be measured. (166)

But how to measure degrees of belief? Ramsey says that there are two possibilities. The first, which he dismisses, is that "the degree of a belief is something perceptible by its owner; for instance that beliefs differ in the intensity of a feeling by which they are accompanied." He argues instead for the second possibility: "that the degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it." This is the idea – a *functionalist* view of degree of belief, as we would now call it – that he then proceeds to develop with characteristic alacrity.

In taking this course, Ramsey is guided by what he calls "the old-established way of measuring a person's belief," which is "to propose a bet and see what are the lowest odds which he will accept." Ramsey finds this method to be "fundamentally sound" (barring some deficiencies due to features like diminishing marginal utility of money, agent's possible disdain for gambling, etc., which can nonetheless be dealt with by stipulating a series of postulates in the formal model).

More precisely, Ramsey considers an agent who chooses among *gambles* of the form

$$\alpha \text{ if } p, \beta \text{ if } \neg p.$$

where p is a proposition and α, β are "goods" that the agent values. The gamble is understood in the usual sense: in accepting this gamble the agent gets α if p is true, β otherwise. For instance, let p be "The result of next toss of this coin is head" and α and β be some monetary rewards (or punishments). An agent who accepts this gamble gets α if the coin lands head, β otherwise.

For notational convenience, let us write $G(p, \alpha, \beta)$ for the gamble that pays α if p, β if $\neg p$. In Ramsey's approach, we think of life as continually presenting us with options of this kind. As he puts it, his model

> is based fundamentally on betting, but this will not seem unreasonable when it is seen that all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. (183)

are a species of probability, so long as the agent concerned satisfies certain coherence constraints.

Agents are thought of as choosing to accept or reject a given bet, or choosing among different bets, based on their *degree of belief* or *credence* in p and the *utility* they assign to α and β.

The agent is assumed to have preferences among gambles of this form. Then, provided that the preference relation among gambles satisfies a set of coherence axioms, the system yields a unique probability function P and a utility function U (unique up to a positive linear transformation) such that the "ultimate good" of accepting gamble $G(p, \alpha, \beta)$ can be represented by expected utilities, that is,

$$EU\big[G(p, \alpha, \beta)\big] = P(p)U(\alpha) + \big(1 - P(p)\big)U(\beta).$$

Ramsey's theory marks the beginning of a long and fruitful development of Bayesian subjectivism, by generations of writers. Philosophically, this approach is motivated by the pragmatic thesis that probability is to be understood in terms of the rational decision making that we, *qua* real-world agents, strive to achieve on a day-to-day basis. Methodologically, it retains at its core Ramsey's operationalized model of personal probabilities (credences) and utilities. Mathematically, it is built with rigorous representation theorems by means of which probabilities can be numerically defined and derived.

For our purposes, what matters is that Ramsey's model gives us an account of *what it is* to hold a credence, or degree of confidence, in a particular proposition. The answer, holistically generated across a space of propositions, in harmony with a simultaneous definition of utility, consists in a disposition to choose certain gambles in preference to others.

3. Can Ramsey's Model Make Sense of Action-credences?

Ramsey says at one point that the agent is assumed to have "certain opinions about all propositions" (174). This cannot be quite correct, however, for there is an important class of propositions to which his model cannot assign non-trivial credences. To see this, consider an agent whose current options include the following gamble:

$$A = G(p, \alpha, \beta).$$

What would it take, in Ramsey's system for this agent to have a credence in whether she will accept A, as she decides whether or not to do so? The answer is that the agent would need to include, in her ranked suite of possible actions, gambles of the form:

$$B = G(\text{I accept } A, \gamma, \delta). \tag{1}$$

For this is the kind of gamble that is relevant to determining whether she has some particular degree of belief in the proposition that she will accept A.

A gamble of the form of B is quite unproblematic if it is considered as a measure of the agent's degree of belief about whether she accepts A *on some other occasion* (a future occasion, or even a past occasion, if we allow that the agent may have forgotten

whether she accepted B at some point in the past). But B makes no sense – or at least, no sense *as a measure of credence – as she decides whether to accept* A. We offer two arguments for this conclusion; we give them as informal arguments here, and in a more formal manner in Appendix A.

3.1. Gambling on one's own choices?

The first argument is that in a context in which an agent is considering gamble A, offering her B simply adds to whatever is already at stake a fixed amount of γ or δ, available to the agent for certain, depending on whether she accepts A. This may give us *some* information about the agent's psychological state – in the limit, as γ and δ are allowed to be large enough to dominate other considerations, it certainly tells us whether she prefers γ to δ, or vice versa (and therefore deprives us of the information that the choice would otherwise provide about other matters). But it tells us nothing about any credence on the agent's part about whether she will do A, as she makes her choice (see Appendix A.1 below for formal details).

We could put the point like this. At the heart of Ramsey's model is a (formalised) notion of *choice.* Agents are assumed to have unrestricted access to a range of options – a range of gambles available to them, each of which they may either accept or decline. Think of this as like a bank of toggle switches: for each switch, the agent is free to set it either on or off. The beauty of the model is to choose the gambles so that the resulting pattern of switch settings reveals the agent's credences over a range of propositions.

Beautiful as it is, this machinery cannot make sense of an assignment of a credence concerning one of switch settings. The chosen switch settings are the 'observables' of the model, on view to the agent concerned as much as to a third party. Until they are fixed, the entire model tells us nothing about the theoretical variables it takes to be underneath (i.e., the agent's credences and utilities); but once they are fixed, there is no room in the model for uncertainty *about them.*

As we saw, the attempt to add a new switch representing a gamble conditional on one of the existing switch settings simply becomes a new reward for the choice of that switch setting. In these circumstances the agent's choice tells us something about their preferences,[3] but nothing new *of an epistemic nature.* Where agents make their own truth, choices that would otherwise reflect degrees of uncertainty have no such significance – and there is simply no substitute, within Ramsey's model.

It is clear that Ramsey recognised this distinction between epistemic matters, on the one hand, and practical matters – things that are up to us – on the other. In a later piece he says this, for example:

> When we deliberate about a possible action, we ask ourselves what will happen if we do this or that. If we give a definite answer of the form

3. Actually not even that, if the gambles are formulated in terms of *goods* – i.e., payoffs whose rankings to the agent are already assumed to be known.

> 'If I do p, q will result,' this can properly be regarded as a material implication or disjunction 'Either not-p or q.' But it differs, of course, from any ordinary disjunction in that one of its members is not something of which we are trying to *discover* the truth, but something it is within our power to *make* true or false. (Ramsey, 1929, 142, emphasis added)

A few lines later, Ramsey again emphasises the non-epistemic character of our relation to propositions concerning our present options:

> Besides definite answers 'If p, q will result', we often get ones 'If p, q might result' or 'q would probably result'. Here the degree of probability is clearly not a degree of belief in 'Not-p or q', but a degree of belief in q given p, which it is evidently possible to have without a definite degree of belief in p, *p not being an intellectual problem.* (142, emphasis added)

Finally, Ramsey also notes that the lacuna in the agent's credences concerns only her present actions – matters currently 'up for decision', as we might say. As his footnote puts it:

> It is possible to take one's future voluntary action as an intellectual problem: 'Shall I be able to keep it up?' But only by dissociating one's future self. (142)

3.2. Self-referential gambles

The point above was that the new gamble B – a bet on whether the agent will accept gamble A – doesn't do the desired job of eliciting a credence about whether the agent will accept A. But there's a deeper concern. It is easy to show – we do so formally in Appendix A.2 below – that adding a gamble B of the form of (1) has the effect of making the set of gambles on offer to the agent *self-referential*, in the sense that she is effectively being offered a gamble on whether she will accept *that very same gamble.* That is, it amounts to introducing a gamble of the form:

$$B' = G(\text{I accept } B', \gamma, \delta).$$

To get a sense of the implications of this without explicit self-reference, suppose that B and A are two of N gambles on offer to the agent. We can combine these N gambles into a single gamble with 2^N options, in the obvious way. Any choice among these options fixes at the A component the very matter on which the choice of the B component attempts to assign a credence.

The effect is that two roles in Ramsey's model are being confused. We are attempting to equate one of the binary-valued *observables* with one of the probability-valued *theoretical terms*.[4] This is mathematically incoherent unless the term in question takes

4. That is, a term x such that $0 \leq x \leq 1$. The attempt to equate these two terms is explicit in the self-referential case. In the informal case just given, a probability-valued theoretical term is being treated as a component of a single 2^N-ary binary-valued observation term.

only the binary values 0 or 1. And it is conceptually incoherent – a kind of category mistake – unless we think of these trivial values as representing simply whether the agent declines or accepts the bet in question (thus leaving credence and belief out of the picture, in this special kind of case). There is simply no place in Ramsey's model for a credence in the acceptance of a self-referential gamble, except in these trivial senses – and as our argument in Appendix A.2 shows, any attempt to make sense of credence in the acceptance of a current gamble is equivalent to the self-referential case.

To summarise, we have shown that there is no room in Ramsey's model for credences for currently-contemplated gambles, or actions. In other words, DCOP holds, *within Ramsey's model,* and *subject to the restriction to present actions.* So, success of a kind for DCOP, but both qualifications are essential. We have just seen that Ramsey himself allows 'remote' action credences; and, as we shall shortly explain, it is a trivial matter to extend Ramsey's model so that it allows present action credences – i.e., so that DCOP fails completely.

3.3. Two connections

Before moving on, we want to note two connections between our exposition of the incoherence of action-credence in Ramsey's model, and familiar points in the existing literature on agency and decision. The first point, which we might characterise in Ramsey's terms as the idea that *making true* supplants *discovering true,* has a very close affinity to themes in the work of David Velleman and Jenann Ismael, among others. It is at the core doctrine of Velleman (1989) that agents enjoy 'epistemic freedom' with respect to their own actions, and of the observation by Ismael (2012) that choices are self-validating epistemic 'wild cards'. Though they express the idea in slightly different ways, Ramsey, Velleman and Ismael seem to have a common intuition in mind: it is of the essence of choice that it involves a kind of *epistemic singularity*, a place in which the rules do not apply in the normal way. We will return to this thought below, and explore its development by James Joyce.

The second point relates to the failure of the gamble-based operational characterisation of credence, in the case of act-credences. In this case our observation is very much in the spirit of several writers who have noted that it seems difficult to make sense of betting on one's own actions, in the context of choice.[5] Again, we make no claim of novelty for our presentation of the point, unless on the grounds that we add something to the explanation of the inapplicability of Ramsey's credence-eliciting

5. Spohn (1977) and Levi (1989) argued that the standard betting interpretation of probability collapses when it is applied to action-events. Their arguments, which involve revisions of rewards and events in a bet, have generated heated debates regarding, among other things, what is the "correct" way to apply the betting interpretation (cf. exchanges on this matter from Levi (2000); Joyce (2002); Rabinowicz (2002); Levi (2007), Spohn (2012)). Gaifman (1999) provided an analysis of self-reference and cyclic reasonings involved in decision and game theoretic models and pointed out that, within the classical Bayesian subjective decision/probability theories, act-credences turn on certain conceptual circularities that cannot be rationally justified and hence should be barred from this framework.

machinery in this special case (especially by pointing out the unavoidability of self-referential gambles if we attempt to add action credences to Ramsey's model).

In one respect, indeed, our claims may be explicitly less ambitious than those of these other authors. For we are explicit that we intend our conclusions to be model-relative, rather than absolute. We have been concerned to explain why DCOP holds, why action-credences cannot exist, *within Ramsey's model.*

4. Beyond Ramsey's Model

As we noted, the reason we have identified for the incoherence of (present) action credences within Ramsey's model is related to the objection, often offered in favour of DCOP, that there is something deeply problematic about offering an agent bets on her own actions as a means of eliciting action credences. This objection is discussed in a classic paper by Rabinowicz (2002), himself an opponent of DCOP. Rabinowicz concedes that there is some merit to the argument, but suggests that it doesn't establish as much as the proponents of DCOP require:

> In those cases when bet offers themselves would influence our probabilities for the events on which the bets are made, probabilities no longer are translatable into betting dispositions. This does not mean, however, that probability estimates are impossible to make in cases like this. The correct conclusion is rather that the connection between probabilities and betting rates is not as tight as one might initially be tempted to think. (Rabinowicz, 2002, 110)

We agree with Rabinowicz, if we interpret him simply as noting the possibility of models that extend the Ramsey framework, to allow the existence of action credences in circumstances in which the original framework does not.[6]

Indeed, the point is a rather obvious one, for here is a simple way to construct such an extension. Imagine that our agent carries in his pocket a Personal Digital Assistant, Siri, who attempts to maintain a dynamic assignment of probabilities to a range of the agent's possible future actions.[7] Now let our hybrid model use the agent's own credences, where available, and Siri's, where not. This hybrid model can certainly assign credences to the agent's presently-contemplated actions – credences originating in Siri – even though the agent's own Ramseyian model cannot.

This somewhat trivial example makes a serious point. Anyone in these debates who, unlike us, takes themselves to be discussing DCOP as a thesis about agents

6. We suspect that Rabinowicz has in mind something stronger, namely that such extensions might be 'more realistic', or otherwise preferable, but we set aside that difference for now.

7. Siri does this in an attempt to keep her agent out of trouble, and is able to do it effectively because she has access to the traditional sources of evidence – the agent's entire history of 'Likes' and 'Dislikes' on Facebook, dinner reservations on Opentable, purchasing history on Amazon, and so on. One of the challenges for proponents of DCOP is to explain why the agent himself cannot use such information to generate credences about his own choices, as he makes them – more on this below.

simpliciter, rather than as about agents *modelled in some particular way*, would do well to ask themselves what they mean by an 'agent'. Without a well-motivated restriction of the field, it risks being true rather trivially that some agents do satisfy DCOP and others do not, so that there is no general thesis to be had.

Are there non-trivial reasons for entertaining models that modify Ramsey's framework so as to admit action credences? Certainly, and to illustrate the point we shall now turn to two motivations offered by James Joyce. Joyce is a particularly interesting case, from our point of view. He is a strong advocate of DCOP-violating models, but he makes moves within them that have much in common with some of the key insights of those who favour DCOP. This offers the interesting prospect that we might be able to identify an important generic feature of agency, common to DCOP-respecting and DCOP-violating models, and itself of considerably more interest than the choice *between* such models. (As we shall see, this possibility turns on the fact that Joyce is not so much *extending* Ramsey's model, as in the Siri case, but *relabelling* it, by treating as 'belief-like' some elements that are present in a Ramseyian model, but classified in a different way.)

4.1. Joyce on the role of action credences

Joyce (2002) criticises two claims made by Isaac Levi: first, as Joyce puts it, the claim "that the causal theory requires deliberating agents to make predictions about their own actions"; and second, the claim "that this is incoherent because 'deliberation crowds out prediction.'" (69)

Concerning Levi's first claim, Joyce argues that "nothing in causal decision theory *forces* an agent to make predictions about her own acts", adding that so far as he knows, he himself is "the *only* causal decision theorist doctrinally committed to rejecting the 'deliberation crowds out prediction' thesis." (69) Concerning the second, he defends the following conclusions:

> [T]he ability of a decision maker to adopt beliefs about her own acts during deliberation is *essential* to any plausible account of human agency and freedom. While Levi suggests that a deliberating agent cannot see herself as free with respect to acts she tries to predict, precisely the reverse is true. Though they play no part in the *rationalization* of actions, such beliefs to are essential to the agent's understanding of the *causal genesis* of her behavior. (70)

Thus Joyce presents himself as a staunch opponent of DCOP – making it all the more interesting when, as we shall see below, he turns out to agree with Ramsey on central points about agency.[8]

When Joyce turns to DCOP, he provides a useful summary of, as he puts it, "some general worries that one might have about letting agents assign probabilities to their own acts":

8. We shall also see that his disagreement with Levi turns mainly on an equivocation about the term 'predict'.

> *Worry-1:* Allowing act probabilities might make it permissible for agents to use the fact that they are likely (or unlikely) to perform an act as a *reason* for performing it.
> *Worry-2:* Allowing act probabilities might destroy the distinction between acts and states that is central to most decision theories.
> *Worry-3:* Allowing act probabilities âĂIJmultiplies entities needlesslyâĂİ by introducing quantities that play no role in decision making. (79)

In each case, he expresses sympathy for the concern, but argues that DCOP is not required in order to meet it. Thus:

> As to Worry-1, I entirely agree that it is absurd for an agent's views about the advisability of performing any act to depend on how likely she takes that act to be. Reasoning of the form "I am likely (unlikely) to A, so I should A" is always fallacious. While one might be tempted to forestall it by banishing act probabilities altogether, this is unnecessary. We run no risk of sanctioning fallacious reasoning as long as A's probability does not figure into the calculation of its own expected utility, or that of any other act. (79–80)

Note that Ramsey's model does not allow anything other than "banishing act probabilities altogether", as Joyce puts it here. In effect, Joyce is agreeing that self-referential gambles are incoherent, and proposing to understand 'credence' so that act credences do not commit us to such gambles. But for Ramsey there is no such option – credences are *defined* in terms of gambles, and in the case of act credences that would require the kind of self-referential gambles that Joyce agrees to be absurd. (It is not clear that the absurdity Joyce has in mind is precisely the one we identified in §3 – arguably, it can't be, for Joyce takes it to obtain even in his non-Ramseyian models. We return to this issue in §5.)

Similarly, concerning Worry-2, Joyce concludes:

> Even if act probabilities do not figure into the calculation of act utilities, they may have other roles to play in the process of rational decision making. Indeed, we shall soon see that they do. (80)

Again, concerning a presentation by Levi of the apparent unmeasurability of action credences within SDT, Joyce says:

> It is quite true that the probabilities [an agent] assigns to acts during her deliberations cannot be elicited using wagers in the usual way, but this does not show that they are incoherent, only that they are difficult to measure. (86–87)

Finally, Joyce makes similar remarks about Worry-3:

> As Wolfgang Spohn has long argued, there is no reason to allow act probabilities in decision theory if we cannot find anything useful for them to do. Given that they play no role in the evaluation or justification of acts, it would seem that there is nothing useful for them to do. Why not abolish them? (98)

Joyce responds as follows:

> Act probabilities are a kind of epiphenomena in decision theory. Though they do no real explanatory work, they are tied to things that do. We need act probabilities because (i) we need unconditional subjective probabilities for *decisions about acts* to *causally* explain action (though not to rationalize it), and (ii) we need Efficacy to explain what it is for an agent to regard acts as being under her control. Efficacy requires that P(A\dA) = P(¬A\d¬A) = 1, and so P(A/dA) = P(¬A/d¬A) = 1. One cannot have these latter conditional probabilities and unconditional probabilities for dA and d¬A without also having unconditional probabilities for A and ¬A. Act probabilities are not only coherent, they are *compulsory* if we are to adequately explain rational agency. We cannot outlaw them without jettisoning other subjective probabilities that are essential ingredients in the causal processes that result in deliberate actions. When it comes to beliefs about one's own actions, deliberation does not "crowd out" prediction; it mandates it! (98–99)

This will take a little unpacking. First, it is important to note that Joyce is distinguishing between an agent's decision to do A, written dA, and the act A itself. When he talks of act probabilities, he means P(A) and P(¬A), not P(dA) and P(d¬A). This is another potential source of talking at cross purposes – some proponents of DCOP may take it for granted that the important issue concerns the latter credences, not the former.

Fortunately this distinction doesn't matter much in this context, because Joyce is equally committed to the need for unconditional probabilities of both kinds, P(A) and P(dA). They are connected by the principle Efficacy, which Joyce takes to encode the idea that the agent takes A to be under her control – if she chooses dA then A results, and similarly for d¬A and ¬A.

Let us grant Joyce that, as he puts it here, "Efficacy requires that P(A\dA) = P(¬A\d¬A) = 1, and so P(A/dA) = P(¬A/d¬A) = 1."[9] It certainly follows that if we allow unconditional probabilities P(dA) and P(d¬A) then we shall have to allow P(A) and P(¬A), as well. But why do these conditional claims require unconditional probabilities in the first place?

We see two possible answers at this point. The first, which we are not sure whether to attribute to Joyce, is that the conditional probabilities are defined in terms

9. The backslash in 'P(A\dA)' represents what Joyce calls ''causal probability' – "it represents [the agent's] beliefs about what her acts will *causally promote*, so that P(S\A) will exceed P(S\¬A) only if [the agent] believes that A will causally promote S." (79)

of unconditional probabilities, so that we cannot have P(A/dA) and P(¬A/d¬A) without having P(dA) and P(d¬A) as well. This is a very familiar move, and needs to be mentioned as a motivation for extending Ramseyian SDT to add unconditional action credences. But it also admits a well-known reply, namely, that there are other reasons for treating conditional probability as primitive, and not defining it as the usual ratio of unconditional probabilities. Ramsey himself favoured this approach. In a passage we quoted above, he refers to one's "degree of belief in *q* given *p*, which it is evidently possible to have without a definite degree of belief in *p*" (1929, 142, and see also his 1926, 180). Later proponents include Renyi (1970), Price (1986b), Mellor (1993), and Hájek (2003).

Whether or not Joyce has this consideration in mind, his main point is a different one. He proposes a detailed model of the deliberative process in which the credences P(dA) and P(d¬A) play a crucial role. As he notes, the model owes much to Velleman. From Joyce's point of view the attractions of this model provide the strongest case for accepting P(dA) and P(d¬A), and hence for rejecting DCOP. Accordingly, we want to follow Joyce's explication of the model in some detail. It is crucial to our claim that (apparent) disagreements about DCOP are obscuring deeper agreement about the nature of agency.

4.2. Joyce on "evidential autonomy"

Joyce introduces his discussion of the model in question by articulating yet another concern about action credences:

> I am portraying the agent who changes her mind as altering her *beliefs* about what she will decide *on the basis of no evidence whatever.* She goes from being certain that she has decided on ¬A to being certain that she has decided on A without *learning* anything. Can this sort of belief change be rational? By letting agents assign subjective probabilities to their own acts it seems that we are also letting them believe whatever they want about them. This means that act probabilities must be radically unlike other probabilities in that they seem not to be at all constrained by the believer's evidence. ... This, I suspect, gets us to what is really bothering people about act probabilities. (2002, 94–95)

In response, Joyce notes first that when an agent

> sees herself as a free agent in the matter of A, Efficacy ensures that all of her evidence about A comes by way of *evidence about her decisions.* Her justification for claiming that she will do A will always have the form: "here is such-and-such evidence that I will decide on A, and (via Efficacy) deciding on it will cause me to do it." (95)

As Joyce says, this may seem "to push the problem back from beliefs about acts to beliefs about decisions." But he argues that "this is not so":

> An agent's beliefs about her own decisions have a property that most other beliefs lack: under the right conditions they are *self-fulfilling*, so that if the agent has them then they are true. Understanding this is one of the keys to understanding human agency and freedom. (95)

Joyce explains this point with reference to Velleman's notion of epistemic freedom:

> According to Velleman, ... the believer has a kind of "epistemic freedom" with respect to self-fulfilling beliefs that she lacks for her other opinions; she can *justifiably* believe whatever she wants about them. If she is sure that believing H will make H true and that believing ¬H will make ¬H true then, *no matter what other evidence she might possess*, she is at liberty to believe either H or ¬H because she knows that *whatever* opinion she adopts will be warranted by the evidence she will acquire *as a result of adopting it*. More generally, any increase or decrease in her confidence in H provides her with evidence in favor of that increase or decrease – the stronger a self-fulfilling belief is, the more evidence one has in its favor. (96)

Like Velleman, Joyce sees this idea of self-fulfilling belief as crucial to a proper understanding of agency: "Velleman holds, as I do, that agents are epistemically free with respect to their own decisions and intentions. ... [T]he idea that agents are epistemically free regarding their own decisions is important and entirely correct." (96–97)

Finally, Joyce applies these ideas to offer a model of the *dynamics* of deliberation:

> During the course of her deliberations [an agent's] confidence in "I decide to do A" will wax or wane *in response to information about A's desirability relative to her other options* (e.g., information about expected utilities). If A and ¬A seem equally desirable at some point in the process, then she will be equally confident of dA and d¬A at that time. If further deliberation leads her to see A as the better option, then her confidence in dA will increase as her confidence in d¬A decreases. These deliberations will ordinarily cease when [the agent] is certain of either dA or d¬A, at which point she will have made her decision about whether or not to perform A *by making up her mind what to believe about dA*. (97)

Joyce notes that while

> this process would be nothing more than an exercise in wishful thinking if [the agent's] beliefs about dA and d¬A were not self-fulfilling, the fact that they are ensures that her subjective probability for each proposition increases or decreases in proportion to the evidence she has in its favor. (97)

He concludes: "This explains how [the agent's] beliefs about what she will decide can be both responsive to her preferences and warranted by her evidence at each moment of her deliberations." (97)

Joyce returns to these ideas in a later piece (Joyce, 2007), and links them to a point made by writers on both sides of debates between causal and evidential decision theory:

> [M]any decision theorists (both evidential and causal) have suggested that free agents can legitimately ignore evidence about their own acts. Judea Pearl (a causalist) has written that while "evidential decision theory preaches that one should never ignore genuine statistical evidence ... [but] actions – by their very definition – render such evidence irrelevant to the decision at hand, for actions change the probabilities that acts normally obey." (2000, p. 109)[10] Pearl took this point to be so important that he rendered it in verse:
>
> > Whatever evidence an act might provide
> > On facts that precede the act,
> > Should never be used to help one decide
> > On whether to choose that same act. (2000, p. 109)
>
> Huw Price (an evidentialist) has expressed similar sentiments: "From the agent's point of view contemplated actions are always considered to be *sui generis*, uncaused by external factors ... This amounts to the view that free actions are treated as probabilistically independent of everything except their effects." (1993, p. 261) A view somewhat similar to Price's can be found in Hitchcock (1996).
>
> These claims are basically right: a rational agent, while in the midst of her deliberations, is in a position to legitimately ignore any evidence she might possess about what she is likely to do. ... A deliberating agent who regards herself as free need not proportion her beliefs about her own acts to the antecedent evidence that she has for thinking that she will perform them. Let's call this the *evidential autonomy thesis.* (Joyce, 2007, 556–557)

Joyce adds a footnote at this point:

> It is important to understand that this freedom only extends to propositions that describe actions about which the agent is currently deliberating, and whose performance she sees as being exclusively a matter of the outcome of her decision. It does not, for example, apply to acts that will be the result of future deliberations. (557)

10. Pearl makes a terminological distinction between 'action' and 'act.' He remarks that "[an] act is viewed from the outside, and action from the inside. Therefore, an act can be predicted and can serve as evidence for the actor's stimuli and motivations .... Actions, in contrast, can neither be predicted nor provide evidence since (by definition) they are pending deliberation and turn into *acts* once executed." (2009, 108)

4.3. Comparing Joyce and Ramsey

At this point, we hope it is clear that there are very deep similarities between Joyce's model and Ramsey's. Ramsey's model does not admit credences for dA and d¬A, though with precisely the same qualification articulated in the footnote from Joyce just quoted: the restriction only applies in the context of current deliberations. For Ramsey the rejection of such credences seems a conceptual matter, as well as a consequence of his operational account of credence. For Ramsey, the truth of dA and d¬A is simply "not ... an intellectual problem," as he puts it – "not something of which we are trying to *discover* the truth, but something it is within our power to *make* true or false." (1929, 142, emphasis added)

In fact, however, only a hair's breadth separates this view from Joyce's. Joyce, too, agrees that the truth of dA and d¬A is not an intellectual problem *of the normal sort,* and that these propositions are within our power to make true or false. He simply represents this special status in a different way. For Joyce, there is a belief during deliberation, albeit one with a special epistemic status (because it is self-fulfilling). For Ramsey there is no belief until it is licensed by the formation of an *intention,* either to dA or d¬A – in other words, until *after* deliberation. But this is little more than a stylistic preference, at least compared to the points of agreement.

In particular, the (apparent) disagreement between Ramsey and Joyce reflects a difference about the use of the term 'belief'. Ramsey is taking for granted what we might call an *epistemically-grounded* conception of belief (and hence of partial belief, or credence). On this conception, as in the standard Bayesian picture, beliefs and credences are only acquired, changed, or updated in the light of new evidence – it is a conceptual truth about beliefs that they are responsive to evidence in this way. One common correlate of this idea is the thesis that beliefs have 'world-to-mind direction of fit' (see, e.g., Humberstone 1992). Nothing counts as a belief unless, in some appropriately normative sense, it is 'trying' to match the world.

Intentions or volitions don't fit this pattern. As Anscombe (1957) famously pointed out, intentions have mind-to-world direction of fit. For Ramsey, then, intentions don't count as beliefs, or partial beliefs, or credences. Ramsey will allow that we have beliefs about our own actions, of course, but they are *downstream* of intentions. When one forms the intention to A, one thereby comes to believe that one will A. For example, one thereby becomes disposed to answer "Yes" to the question "Will you A?" But the intention itself is not a belief, according to this epistemically-grounded conception of belief.

In contrast, Joyce, following Velleman, thinks of the intentions we form when we deliberate *as* beliefs – reflexive beliefs about what we ourselves will do. This is why beliefs about one's own action are, as Joyce says, "essential" to his model of deliberation. In Joyce's model the *products* of the process of deliberation – gradually-strengthening intentions to *do* something – just *are* such beliefs. Unlike other beliefs, however, these particular beliefs are self-fulfilling and not responsive to evidence of the usual sort – that's what EAT tells us.

There are two ways to get to Joyce's view from Ramsey's. One is to modify the

epistemically-grounded conception of belief to allow a special class of exceptions, a special class of beliefs whose genesis does not require evidence – namely, the self-justifying beliefs that Joyce identifies with intentions. The other is to stretch the notion of evidence just enough to allow that these special beliefs are supported by evidence after all – self-supported, in effect, by the evidence that they themselves generate or constitute. Whichever way we stretch our terminology, the principle that beliefs have world-to-mind direction of fit gets a little bit stretched, too, but again with the reassurance that these are special cases. Joyce himself is clear that they are special cases. As he remarks: "act probabilities must be radically unlike other probabilities." (2002, 94)

By way of comparison, here is Jenann Ismael's negotiation of the same terminological boundary, with Wittgenstein in Ramsey's shoes and Ismael herself in Joyce's:[11]

> Wittgenstein ... thinks that for [one's own intentions] to count as knowledge, they would have to be subject to the game of certainty and doubt, and that it would have to make sense to doubt their truth. And so for him, these cannot count as genuine knowledge. On the performative model,[12] they are still knowledge, but degenerate because self-fulfilling. Whereas Wittgenstein is suspicious of the idea of knowledge free of epistemic constraints, the performative model explains it and uses it to understand how it shapes the first-person/third-person asymmetries in predictive opinion. Both of us agree that it is wrong to see the sort of certainty we have about our own beliefs on the model of Cartesian transparency based in an introspective faculty. But the performative model provides an alternative that secures the special epistemic status and integrates it neatly with other truth-bearing discourse without undermining its status as knowledge. (Ismael, 2012, 158–159)

In this case, as for Ramsey and Joyce, it is clear that the two views in question are extremely close, easily mapped from one to the other with small variations in terminology. Some readers may feel that there is an interesting question whether the Ramsey/Wittgenstein model or the Joyce/Ismael model comes closer to getting the psychology of decision *right*, but for our purposes what matters are the similarities. The crucial point of agreement is that the fact that supports DCOP in Ramsey's model – i.e., that in the process of deliberation we come to beliefs about what we will do *after* but not *before* we form our intention – is mirrored under a different name in Joyce's picture. For Joyce, it is simply EAT itself, which implies that the beliefs about our own actions that play the role of intentions are not themselves evidentially 'downstream' of other beliefs.

11. Ismael notes the similarity between her view and Joyce's: "James Joyce comes to much the same conclusion .... He writes 'an agent's beliefs about her own decisions are *self-fulfilling*, and that this can be used to explain away the seeming paradoxical features of act probabilities.'" (Ismael, 2012, 156)

12. This is Ismael's label for the view that decisions are self-fulfilling beliefs.

Whichever model we choose, deliberation turns out to crowd out something. For Ramsey, as we have seen, it crowds out action credences. For Joyce it crowds out either (a) the requirement that beliefs be epistemically-grounded, or (b) a particular (third person?) conception of evidence, that prevents an agent's decision as counting as evidence for its own truth. But what does the crowding out is the same in all three cases – it is EAT. Having agreed on this much, the alternative models are effectively isomorphic[13] – only their use of the terms 'belief' and 'evidence' differs.[14]

4.4. Beyond EAT

To put this irenic diagnosis in context, we want to note that for EAT, as for DCOP, it is a trivial matter to find models of cognitive systems acting in the world that do not satisfy this principle. Our Siri-enhanced Ramseyian agent again provides an example. In that case, Siri does "proportion her beliefs about her [agent's] acts to the antecedent evidence that she [Siri] has for thinking that [her agent] will perform them" (to paraphrase Joyce's own statement of EAT). So if we think of Siri and her agent as a kind of composite, extended agent, we do get a formal violation of EAT.

Defenders of EAT are likely to reply that such composites do not deserve to be called agents (or not *just* agents – perhaps the addition of Siri produces an extended mind, one submodule of which is properly called an agent). We have considerable sympathy for this viewpoint, but won't try to defend it here. We mention the example for two reasons. First, we want to reiterate our earlier observation that the kind of issues we have been discussing involve a great deal of model-relativity. It is helpful to think about one's terms. But second, we do think it plausible that EAT marks an important boundary, and take ourselves to be agreeing with both Ramsey and Joyce on this point.

However, our main claim is that DCOP as such does *not* mark such a boundary in these matters. To paraphrase Sayre's Law, it may be that the reason that debates about DCOP seem so intractable is that there is nothing of significance at stake.

---

13. At least when viewed from sufficient distance. We are ignoring the details of Joyce's complex model of deliberation here – the model in which act credences play what Joyce takes to be their essential role in producing action as it is detailed in e.g., Joyce (2012) – in order to highlight the broad structural parallel with Ramsey's model.

14. Similar remarks apply to Joyce's 'disagreement' with Levi, in our view. When Joyce says in a passage we quoted above that "[w]hile Levi suggests that a deliberating agent cannot see herself as free with respect to acts she tries to predict, precisely the reverse is true," (2002, 70) his point rests on the facts that he is using 'predict' in the sense that allows intentions to count as beliefs and predictions. However, Levi would not deny (of course) that a free agent can form intentions. But he uses 'predict' in the more restricted Ramseyian sense, and denies that a free agent can makes prediction *in that sense* about her own present actions. And Joyce agrees with that principle – it is EAT, effectively. In other words, terminological disagreements aside, Joyce is simply not denying what Levi asserts.

5. Why EAT?

Where does EAT itself come from? We want to conclude by proposing an answer to this question. If correct, it reinforces our conclusion that Ramsey and Joyce are really on the same page, and shows what must be denied by anyone who wants to disagree. We'll introduce this diagnosis by raising a further puzzle about Joyce's view.

5.1. Queries for Joyce

As we saw, Joyce formulates EAT as follows:

> A deliberating agent who regards herself as free need not proportion her beliefs about her own acts to the antecedent evidence that she has for thinking that she will perform them. (2007, 557)

But *why* is this so, according to Joyce, and precisely *when* is it so? Compare the case of a coin toss. Imagine a coin that says 'The result is Heads' on one side and 'The result is Tails' on the other. Whichever statement turns out to be visible when the coin is tossed is self-justifying, but that doesn't stand in the way of our having evidence about the result in advance, let alone give us grounds to ignore such evidence, at that point. Is deliberation different, according to EAT (and Joyce)? In the coin toss, too, we needn't apportion our beliefs *after* the toss to the *antecedent* evidence, but that isn't news.

If there is to be something distinctive about the case of free action, not present in the coin toss case, EAT needs to apply either *before* the choice, or somehow *during* the choice. The latter possibility seems to make most sense, from Joyce's point of view. Choice is a matter of adopting a belief about what one will do, in Joyce's model. Read this way, EAT tells us that adoption of belief *during the process of choice* isn't constrained by prior evidence. (The coin toss analogy now works in Joyce's favour. The statement on display after the toss is entirely justified, even if the antecedent evidence made it very unlikely.)

But what is Joyce's view about an agent's "beliefs about her own acts" *at the beginning* of the process of choice? Does she take over credences based on antecedent evidence about how she will act, or does EAT already rule that out? (When does the EATing start, as it were?) There may be a clue in Joyce's remark that act-credences cannot be reasons for acting:

> [I]t is absurd for an agent's views about the advisability of performing any act to depend on how likely she takes that act to be. Reasoning of the form "I am likely (unlikely) to A, so I should A" is always fallacious.

On the face of it, this suggests that Joyce allows that an agent can hold act-credences right at the beginning of a deliberation, but thinks that it would be absurd to take credences as reasons for one's choice. But why should that be so? Thinking that I am likely to do A, I choose to do so in order to confirm my own present prediction. That's a somewhat 'self-satisfied' reason, perhaps, but what makes it absurd?

It might seem that the absurdity follows from EAT. In virtue of EAT, the beliefs formed *during* deliberation cannot be evidentially constrained by prior evidence. Once I have chosen to A, my reason for thinking that I will A is that I have formed the self-validating belief that I will A, and prior evidence is irrelevant, at this point.

But this can't be right diagnosis. When Joyce says that "[r]easoning of the form 'I am likely (unlikely) to A, so I should A' is always fallacious," he isn't talking about an *epistemic* fallacy, or a mistaken piece of *evidential* reasoning. Joyce's remark is about reasons *for acting*. Whichever belief I choose (in Joyce's model), it will be self-validating, but presumably I can have *non-epistemic* reasons for choosing one action rather than another. Joyce's claim here is that an action credence can't be a reason *of that non-epistemic sort*.

Perhaps EAT is doing the work indirectly? In virtue of EAT, pre-deliberation action credences are liable to be 'evidentially unstable' – an unreliable guide to future credence on the same matter, as it were. EAT ensures that there is no *epistemic* constraint that requires that post-choice act credences align with pre-choice act credences. So treating a pre-deliberative action credence as a reason would be sitting on a stool one of the legs of which is liable to collapse under your weight – guaranteed to collapse, perhaps, in the sense that EAT ensures that the pre-choice act credence carries no authority whatsoever, after the choice is made.

Indeed, if we were to allow pre-choice act-credences to be reasons they would be liable to undermine themselves *before* the choice ever got to be made. If my pre-choice credence that I will do A feeds into my decision to do A, then in arriving at that credence I acquire new evidence relevant to whether I will do A – I learn of a new reason relevant to my choice. But this is liable to *change* my pre-choice credence. At the very least, it means that there is new evidence for me to consider.[15]

We think that these considerations are moving in the right direction, but that they don't get to the heart of the matter. What we need is an explanation for the fact that pre-choice act credences cannot be attached to the deliberative stool in the first place. We think that such an explanation – indeed, an explanation for EAT itself – can be found in a well-recognised cognitive phenomenon known as 'transparency'. We turn to an insightful account of this phenomenon by Richard Moran.

5.2. Moran on transparency

Moran (2001, §2.6) describes the transparency of first-person present-tensed thought as follows (attributing the term to Roy Edgley):

> Ordinarily, if a person asks himself the question "Do I believe that P?," he will treat this much as he would a corresponding question that does not refer to him at all, namely, the question "Is P true?" And this is not how he will normally relate himself to the question of what someone else

15. Ismael (2012, 160) notes that Jonathan Bennett makes a similar point about the instability of predictions about our own behaviour, while we deliberate. And Price's (1986a, 1991) defence of Evidential Decision Theory relies on a similar instability argument, motivated by the Principle of Total Evidence.

believes. Roy Edgley has called this feature the "transparency" of one's own thinking:

> [M]y own present thinking, in contrast to the thinking of others, is transparent in the sense that I cannot distinguish the question "Do I think that P?" from a question in which there is no essential reference ence to myself or my belief, namely "Is it the case that P?" This does not of course mean that the correct answers to these two questions must be the same; only I cannot distinguish them, for in giving my answer to the question "Do I think that P?" I also give my answer, more or less tentative, to the question "Is it the case that P?" (Edgley, 1969, 90)

We can't do justice here to Moran's careful discussion of this idea, but we'll highlight a few central points.[16] Moran offers the following diagnosis of transparency:

> [T]he claim of transparency is that from within the first-person perspective, I treat the question of my belief about P as equivalent to the question of the truth of P. ... [W]hat ... transparency requires is the deferral of the theoretical question "What do I believe?" to the deliberative question "What am I to believe?" And in the case of the attitude of belief, answering a deliberative question is a matter of determining what is true. (62-3)

This diagnosis involves a distinction between two epistemic stances on one's own mind, a distinction that Moran describes like this:

> In characterizing two sorts of questions one may direct toward one's state of mind, the term 'deliberative' is best seen at this point in contrast to 'theoretical,' the primary point being to mark the difference between that inquiry which terminates in a true description of my state, and one which terminates in the formation or endorsement of an attitude. (63)

Moran argues that the deliberative stance is central and ineliminable in our cognitive lives. Concerning the suggestion that we might somehow dispense with the deliberative inquiry, relying solely on the theoretical inquiry, he responds:

> The problem with the idea of generalizing the theoretical stance toward mental phenomena is that a person cannot treat his mental goings-on as just so much data or evidence about his state of mind all the way down, and still be credited with a mental life (including beliefs, judgments, etc.) to treat as data in the first place. (150)

16. For another insightful discussion of transparency and its relevance to an understanding of the agent's perspective, we recommend Ismael (2012). We ourselves discuss these issues at greater length in [A1 & A2, 2017].

Moreover, Moran takes the lessons of transparency to apply equally to deliberation about what to *avow* and deliberation about what to *do.* As he puts it:

> [W]e might . . . compare the case of belief with that of knowledge of one's own future behavior: a person may have a purely predictive basis for knowing what he will do, but in the normal situation of free action it is on the basis of his decision that he knows what he is about to do. In deciding what to do, his gaze is directed "outward," on the considerations in favor of some course of action, on what he has most reason to do. Thus his stance toward the question, "What am I going to do now?" is transparent to a question about what he is to do, answered by the "outward-looking" consideration of what is good, desirable, or feasible to do. (105)

For action, as for belief, Moran emphasises that transparency does not mean that the agent does not have knowledge of her own state of mind. The point is rather that that knowledge comes from a distinctive source, only available in the first-person present-tensed case – via a *deliberative* path, rather than a *theoretical* or *empirical* path, as Moran puts it. The last passage continues:

> When [the agent] answers this question [i.e., "What am I going to do now?"] for himself and announces what he is going to do, . . . [w]hat he has gained, and what his statement expresses, is straightforward knowledge about a particular person [i.e., himself], knowledge that can be told and thus transferred to another person who needs to know what he will do. (105-6)

Borrowing a term from our own context, we might characterise Moran's conclusion as being that from the first-person present-tensed perspective the deliberative path to knowledge *crowds out* the theoretical path (though the content of the knowledge achieved is precisely the same).

5.3. From transparency to Evidential Autonomy

Let us now apply these ideas to our own discussion. In the section before last we were looking for a justification for the thesis (required by Joyce's claim that "it is absurd for an agent's views about the advisability of performing any act to depend on how likely she takes that act to be") that a pre-deliberative act credence cannot be a reason for the action in question. As we put it there, why can't pre-deliberative act credences support the deliberative stool?

Transparency gives us an answer. In effect, it implies that deliberation turns the stool upside down. It makes our knowledge of what we will do *rest on* the deliberative seat, and not vice versa. What is absurd about taking act credences to be reasons is that during deliberation, deliberation itself is the *source* of one's act credences. At this point, trying to take an act credence to be a reason is simply putting the cart before

the horse – one needs one's reasons in order to *generate* one's act credences.[17]

More generally, Moran's distinction between two paths to knowledge of ourselves – the *theoretical* path and *deliberative* path – offers us a straightforward explanation of EAT itself. EAT simply rests on the fact that when we embark on the deliberative path, we set aside the theoretical path. That's the core of transparency.

Indeed, this diagnosis suggests that Joyce's own formulation of EAT is a little too weak. As Joyce expresses it, EAT is this principle:

> A deliberating agent who regards herself as free *need not* proportion her beliefs about her own acts to the antecedent evidence that she has for thinking that she will perform them. (2007, 557, emphasis added)

This suggests a picture in which the antecedent evidence is still sitting there, as it were, but the agent simply has the option of ignoring it, in deciding to proportion her beliefs about her own acts. Moran's picture is more exclusive, and less voluntary. *By* deliberating, we move ourselves out of the evidential space altogether. In particular, the agent doesn't have the option of *not* ignoring the evidence, because she is no longer playing the evidential game, no longer following the theoretical path.

Once again, various terminological options present themselves at this point. Joyce may prefer to say that the agent doesn't leave evidential space altogether, but rather enters a special kind of evidential space (one in which she calls the evidential shots, so to speak). Again we want to bracket these terminological issues, in order to focus on the underlying structural bifurcation that seems agreed on all sides. This is that deliberation involves a distinctive path to knowledge of our own present choices – a path that takes precedence over, indeed 'crowds out', the theoretical path that we rely on in third person and non-present-tensed cases.

This bifurcation, or separation between two paths to knowledge of our own actions, is what transparency explains. (Indeed, if Moran is right, it is simply the special practical case of something more general.) We propose that it is the source of EAT, and at the heart of what is correct about DCOP.

A. Appendix

Here we present the formal versions of the two arguments given in §3.

A.1. A gamble on another gamble is not a display of credence

Consider, for simplicity, the case where the agent has only two gambles to choose from, namely $A = G(p, \alpha, \beta)$ and $B = G(\text{I accept } A, \gamma, \delta)$ as formulated in §3. Now

17. True, one might take the *memory* of a pre-deliberative credence to provide a reason – I'm doing it because I predicted that I would, and I want to prove myself right. But here the reason is not the prior credence itself, but the belief that one previously held that credence. We are mentioning the credence, not using it, so to speak.

suppose that the agent's credence on proposition p is r_p and, for *reductio*, her credence on 'I accept A' is r_A. Then, given her options, the following are the possible consequences the agent may end up with as results of her actions:

Table 1

act	utility value
$A\&B$	$\gamma + EU(A) = \gamma + [r_p\alpha + (1 - r_p)\beta]$
$\neg A\&B$	δ
$A\&\neg B$	$EU(A) = r_p\alpha + (1 - r_p)\beta$
$\neg A\&\neg B$	0

(where '$A\&\neg B$' reads "I accept A but reject B," and so on.) It is plain that, in this case, the decision problem reduces to a simple choice problem among different consequences of her actions. Then, the agent should just act in a manner that maximizes her gain. But such choices tell us nothing about any credence on the agent's part about whether she will do A – the act-credence r_A has nothing to do with the situation.

A.2. Gambling on gambles leads to self-referential gambles

There is a deeper objection to using gambles of the form of (1) to elicit credences about an agent's own actions, in the context in which those actions are under consideration. This is because any gamble of the form of (1) – gambles whose formulation involves act-propositions that describe the agent's available options – is equivalent to a gamble whose determination contains references to itself.

To see this, let's use our running example in Section A.1 above where the agent is offered two gambles, namely $A = G(p, \alpha, \beta)$ and $B = G(\text{I accept } A, \gamma, \delta)$. As before, the agent has four options: $A\&B$, $A\&\neg B$, $\neg A\&B$, and $\neg A\&\neg B$. Suppose that the agent's credence on "I accept A" is such that $0 < r_A < 1$.[18] Write gamble B in its original form:

$$B = \gamma \text{ if I accept } A; \delta \text{ if I reject } A. \tag{2}$$

Now, given the agent's available options, the case 'I reject A' can be distinguished into two subcases, namely 'I reject A but not B' and 'I reject both A and B'. Similarly, the case 'I accept A' can be further distinguished into two subcases, namely 'I accept only A but not B' and 'I accept both A and B'. Then (2) takes the form under these subdivisions as

$$\begin{aligned} B = &\gamma_1 \text{ if I accept both } A \text{ and } B, \gamma_2 \text{ if I accept } A \text{ but not } B; \\ &\delta_1 \text{ if I reject } A \text{ but not } B, \delta_2 \text{ if I reject both } A \text{ and } B, \end{aligned} \tag{3}$$

18. Unless we are prepared to say that act-credences like r_A are strictly a zero-or-one matter (i.e., $r_A = 0$ or 1), which reduces the current decision problem to choices among different payoffs as presented in the last subsection. Otherwise we are left with the option of assuming the existence of non-trivial credence $0 < r_A < 1$.

where γ_1 and γ_2 are payoffs under actions $A\&B$ and $A\&\neg B$ respectively, and similarly for δ_1 and δ_2 (in our running example, these values accord well with the ones in Table 1.) Further, the action credence r_A, if exists, can be partitioned into $r_{A\wedge B}$ and $r_{A\wedge\neg B}$; and, similarly, $r_{\neg A}(= 1 - r_A)$ into $r_{\neg A\wedge B}$ and $r_{\neg A\wedge\neg B}$.

Notice that the first and the third term of (3) are subcases of 'I accept B', and the second and the fourth term are subcases of 'I reject B'. We can then regroup them and recast (3) as follows huw

$$B = \gamma' \text{ if I accept } B; \delta' \text{ if I reject } B. \tag{4}$$

This yields a *self-referential* gamble $B = G(\text{I accept } B, \gamma', \delta')$, whose conditions of determination refer to its own acceptance or rejection. And this gamble is to be represented in the current model by

$$EU(B) = \gamma' r_B + \delta' r_{\neg B} \tag{5}$$

where $r_B = r_{A\wedge B} + r_{\neg A\wedge B}$ and $r_{\neg B} = r_{A\wedge\neg B} + r_{\neg A\wedge\neg B}$.

It is easy to see that the above argument can be generalized to be applied to cases with more than two options. The argument shows that from *any* gamble of the form of (2) one can always recover a self-referential gamble of the form of (4). In other words, any gamble like $B = G(\text{I accept } A, \gamma, \delta)$ is a self-referential gamble in disguise!

References

Ahmed, A. (2014). *Evidence, Decision and Causality*. Cambridge University Press.

Anscombe, G. E. M. (1957). *Intention*. Harvard University Press.

Bernoulli, J. (1713). *Ars conjectandi*. Impensis Thurnisiorum, fratrum.

Borel, É. (1924). Apropos of a treatise on probability. *Studies in Subjective Probability*, pages 46–60.

Edgley, R. (1969). *Reason in Theory and Practice*. London: Hutchinson.

Gaifman, H. (1999). Self-reference and the acyclicity of rational choice. *Annals of Pure and Applied Logic*, 96(1-3):117 – 140.

Hájek, A. (2003). What conditional probability could not be. *Synthese*, 137:273 – 323.

Hájek, A. (2016). Deliberation welcomes prediction. *Episteme*, 13(4):507–528.

Hitchcock, C. R. (1996). Causal decision theory and decision-theoretic causation. *Nous*, 30(4):508–526.

Humberstone, I. L. (1992). Direction of fit. *Mind*, 101(401):59–83.

Ismael, J. (2012). Decision and the open future. In Bardon, A., editor, *The Future of the Philosophy of Time*, pages 149–168. Routledge.

Joyce, J. (2002). Levi on causal decision theory and the possibility of predicting one's own actions. *Philosophical Studies*, 110(1):69–102.

Joyce, J. M. (2007). Are newcomb problems really decisions? *Synthese*, 156(3):537–562.

Joyce, J. M. (2012). Regret and instability in causal decision theory. *Synthese*, 187(1):123–145.

Laplace, P.-S. (1810). Analytic theory of probabilities. *Paris: Imprimerie Royale*.

Levi, I. (1989). Rationality, prediction, and autonomous choice. In *The Covenant of Reason: rationality and the commitments of thought*, pages 19–39. Cambridge University Press 1997.

Levi, I. (1996). Prediction, deliberation and correlated equilibrium. In *The Covenant of Reason : rationality and the commitments of thought*, chapter 5. Cambridge University Press. 1997.

Levi, I. (2000). Review essay: The foundations of causal decision theory. *The Journal of Philosophy*, 97(7):387–402.

Levi, I. (2007). Deliberation does crowd out predecition. In Ronnow-Rasmussen, T., Petersson, B., Josefsson, J., and Egonssson, D., editors, *Homage a Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz. E*.

Mellor, D. H. (1993). How to believe a conditional. *The Journal of Philosophy*, 90(5):233–248.

Moran, R. (2001). *Authority and Estrangement: an essay on self-knowledge*. Princeton University Press.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge University Press.

Pearl, J. (2009). *Causality: models, reasoning and inference*. Cambridge University Press, 2 edition.

Price, H. (1986a). Against causal decision theory. *Synthese*, 67(2):195–212.

Price, H. (1986b). Conditional credence. *Mind*, 95(377):18–36.

Price, H. (1991). Agency and probabilistic causality. *The British Journal for the Philosophy of Science*, 42(2):157–176.

Price, H. (1993). The direction of causation: Ramsey's ultimate contingency. In *Proceedings of the Biennial Meeting of the Philosophy of Science Association 1992*, pages 253–267.

Rabinowicz, W. (2002). Does practical deliberation crowd out self-prediction? *Erkenntnis*, 57(1):91–122.

Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B., editor, *The Foundations of Mathematics and other Logical Essays*, chapter VII, pages 156–198. London: Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company, 1931.

Ramsey, F. P. (1929). General propositions and causality. In Mellor, D. H., editor, *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, pages 133–51. Routledge and Kegan Paul, 1978.

Renyi, A. (1970). *Foundations of Probability*. Holden-Day, Inc.

Spohn, W. (1977). Where Luce and Krantz do really generalize Savage's decision model. *Erkenntnis*, 11(1):113–134.

Spohn, W. (2012). Reversing 30 years of discussion: Why causal decision theorists should one-box. *Synthese*, 187(1):95–122.

Velleman, J. D. (1989). Epistemic freedom. *Pacific Philosophical Quarterly*, 70:73–97.