Australasian Journal of Philosophy (2019)
doi: 10.1080/00048402.2018.1427119

Heart of DARCness

Yang Liu and Huw Price
University of Cambridge

There is a long-standing disagreement in the philosophy of probability and Bayesian 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? No, say some philosophers, for Deliberation Crowds Out Prediction (DCOP), but others disagree. In this paper, we propose a valid core for DCOP, and identify terminological causes for some of the apparent disputes.

1. Introduction

Can an agent hold a meaningful credence that she will perform an action, as she decides whether to do so? No, say some (e.g., Spohn, 1977; Levi, 1997), for Deliberation Crowds Out Prediction (DCOP). In our view the DCOP debate pays insufficient attention to the meaning of terms such as 'credence', 'deliberation', and 'agent'. We note below that if these terms are understood sufficiently broadly, it is trivial that DCOP fails. Nevertheless, we argue, there are familiar understandings of these terms such that DCOP holds, for reasons related to the so-called 'transparency' of an agent's present-tensed access to certain of her own psychological states. In that sense, we defend DCOP, explaining it in terms of transparency.

One recent critic of DCOP is Alan Hájek (2016), who renames it the DARC Thesis: Deliberation Annihilates Reflexive Credences. We employ Hájek's discussion in three ways. We adopt his terminology, characterising our own project as a quest for the heart of DARCness. We hang our proposal that DARCness rests on transparency on a puzzle that Hájek expresses clearly: Shouldn't an agent be expected to know *more*, not *less*, than other observers about her own mental states? (Transparency explains why we should answer 'no'.) And we build our discussion of the importance of clarifying terminology on examples inspired by another of Hájek's objections to DARC, viz., that it requires implausible 'credal gaps' during deliberation.[1] We describe familiar cases exhibiting analogues of credal gaps – they suggest that this objection to DARC is misplaced, and illustrate the importance of the terminological issues we bring to the fore.

To introduce these examples, we begin with a traveller's tale.

1. Rabinowicz's classic (2002) discussion of DCOP raises a similar objection.

1.1. Bugging Big Al

It's your first time in Benguela, Angola. You're driving to the airport for a meeting with Big Al, who is flying in from the Congo, via Luanda. Confused by your Portuguese-speaking SatNav, you make another wrong turn. "Recalcular", says the SatNav patiently, before displaying a new route. Not for the first time, you wish that Al had more conventional tastes in conversational venues.

Arriving at the airport at last, you check the Arrivals Board. Al's flight, the daily LAD-BUG service, is 45 minutes late. "Don't let the big lad bug you", you think to yourself lamely, looking around the terminal and wondering why Al was so keen to meet in such a dull place. As if reading your thoughts, the Arrivals Board clacks into life, its flaps spinning noisily as it refreshes. Eventually it stops – the flight is now 90 minutes late! Afraid to annoy the Arrivals Board further, and in search of refreshment of your own, you head for the bar. "Cool as a Cuca bar", you say to yourself, ordering the local beer. The television is showing tennis. It's the 2016 Hopman Cup, and Jeff Sock is losing to Lleyton Hewitt – losing with remarkable grace, after encouraging Hewitt to challenge the 'Out' call on one of Hewitt's own serves. Hewitt does so, winning the challenge, the point and eventually the match.

By the time Al's flight arrives, several Cucas later, you've noticed something that all these cases – the SatNav, the Arrivals Board, and the tennis challenge – have in common. In all of them a source of epistemic authority is temporarily *suspended*, as the system updates. In each case there may well be a record of the previous state of the system, before it updates. But there's an obvious sense in which that state has lost its claim to guide our actions, in normal circumstances. Normally, we should just wait for the update to finish, and be guided by the new state of the system.

It occurs to you that this is relevant to your meeting with Big Al. You are here to discuss what Hájek calls the *DARC Thesis*. Hájek opposes DARC, and some of his criticisms turn on the supposed disadvantages of 'credal gaps' – the idea that some propositions cannot be assigned a credence, under certain circumstances. DARC claims that propositions about one's own actions lack a credence, in the context of deliberation. So, as Hájek points out, it commits us to credal gaps.

You know that Big Al and Hájek are as one on these matters. The first thought that strikes you is that updating produces credal gaps in an unremarkable way. The route to the airport, the arrival time, the status of Hewitt's serve – in each case, the epistemic state of the system, and of *you*, if you were taking your guidance from it, was effectively *suspended*, during the updating process. This is bound to happen, for any system that cannot update instantaneously. If a formal model of the epistemic state of such a system, say an assignment of credences to propositions, cannot accommodate such gaps, the fault lies in the model, not in reality.

In honour of Benguela, you give this commonplace kind of credal gap a name. You call it BUG, for *Benign Updating Gap*. The second thought that strikes you is that BUG may be a clue to DARC. If deliberation is a kind of updating then the credal gaps required by DARC may be coming from a familiar place – a suggestion with serious bugging potential, you think, as Big Al joins you in the bar.

2. Credal Gaps

Hájek (2016) points out that DARC commits us to "propositions to which one cannot rationally assign credences" (508). In Hájek's view, these are problematic: "Credence gaps wreak havoc for various foundational notions in probability theory, and thus in formal epistemology." (510) "The stakes are high," as Hájek puts it.

In our view, the proposal that DARC is a BUG suggests that the stakes are actually low. The kind of credal gaps required by DARC are common elsewhere, on any realistic understanding of how we update our beliefs. To illustrate, consider Hájek's core challenge on behalf of "orthodox Bayesianism":[2]

> [T]he DARC Thesis imposes a new constraint on credences beyond the usual Bayesian ones that they obey the probability calculus and are revised by conditionalization. This ... seems to *contradict* orthodox Bayesianism. Consider a case in which initially you idly contemplate a decision that you will make some time in the future; time passes, and eventually the time to make the decision arrives, and you begin deliberating about it; then, finally, you have made your decision. It is perfectly compatible with the DARC Thesis that at the initial and final times, you have credences for what you will do ... Then we have *two* revisions of your probability function: your credences for your options suddenly *vanish* when your deliberation commences, and they suddenly *appear* when it ends. *Neither revision takes place by conditionalization.* This is inconsistent with the Bayesian orthodoxy that credences should be revised by conditionalization (or Jeffrey conditionalization). (511)

But consider a Bayesian agent, Marple, who updates by conditionalization. Let her be a real agent rather than an ideal agent, in that the process of conditionalization takes finite time. We are interested in modelling Marple's credences in a range of propositions – 'The butler did it', say – as she processes new pieces of evidence. Until lunchtime Marple was 75% confident that the butler did it. At lunch she discovered new evidence, which (we are aware) she judges to be relevant. She is presently pacing up and down in the garden, deep in thought. How should we describe her present credal state?

It is easy to see the appeal of adding a description to our model, to capture the distinctive epistemic character of Marple's situation. Marple is *recalculating*, as we might say, remembering the SatNav. She is *between* credal states of the normal sort – the one she held before lunch, and the one she will hold when she returns from the garden. If PC Plod tells us "Miss Marple is 75% confident that the butler did it, Sir, and she is considering new evidence even as we speak," we'll think that he's being even more flat-footed than usual. What he should have said is that she *was* 75% confident, *but* is considering new evidence, recognising that it would be misleading simply to report her credence as 75%. But what other credence could Plod properly give us?

2. As noted, Rabinowicz (2002, 92) makes a similar point.

We don't yet know what Marple will believe when she comes in for tea. What we do know is that her views are presently in flux – she's updating, after all.

If we do say this we have to swallow the consequences. To paraphrase Hájek: "We have *two* revisions of Marple's probability function: her credences suddenly *vanish* (or enter some state of suspension), when her updating commences, and they suddenly *appear* (or get unsuspended), when it ends. *Neither revision takes place by conditionalization.*" Is this result "inconsistent with the Bayesian orthodoxy that credences should be revised by conditionalization"? Yes, if we interpret the Bayesian orthodoxy in a mindlessly literal sense – but so much the worse for that version of the orthodoxy.

In other words, far from representing a problem for the kind of credal gaps associated with DARC, Bayesian conditionalization illustrates the ubiquity of similar credal gaps, in models of how real agents update their beliefs. Bayesian updating itself seems bound to produce credal gaps.

In a moment we qualify this conclusion, noting that if we interpret the term 'credence' sufficiently broadly, it is easy to plug these gaps – hence the importance of terminological issues. But before that, let's note that the Marple case has other features that Hájek takes to create difficulties for DARC.

2.1. More from Marple

Hájek objects that DARC "conflicts with the Reflection Principle," according to which "your credence in *X* at a time should be your expectation of your credences in X at any later time."

> Suppose that now you have not yet begun deliberating about the wine to accompany your dinner, and that you assign a credence to your choosing the red wine that evening – say, 1/2 ... Suppose also that you believe that you will be deliberating about the choice of wine at 7 pm. According to the DARC Thesis, your credence for red wine should be undefined then, so your expectation of your 7 pm credences for red wine should be undefined now. But your credence now is 1/2, not undefined, in violation of Reflection. (511)

But a realistically-modelled Bayesian agent has the same consequence. If Marple knows before lunch that new evidence will present itself at lunch (but nothing yet about where that evidence points), and she knows that she is likely to spend the afternoon updating, then a literal application of Reflection would require that her credence be suspended before lunch, since she expects that it will be so after lunch. So much the worse for a literal reading of Reflection.[3]

Again, Hájek points out that credal gaps associated with action will ramify to the things that agents take to be 'downstream' of those actions:

3. A referee notes that it could be objected that the Reflection Principle applies only to credences that are already defined; but as they also note, the same objection would apply to Hájek's argument.

> I am deliberating between going to Italy and going to Iraq for a holiday, and I believe that my wife will be worried if and only if I go to Iraq. By the DARC Thesis, I cannot have a credence for 'my wife will be worried' – if I did, it would fix my credence for 'I go to Iraq' at the same value. (512)

But similarly here: if Marple believes that if the butler did it then he's likely to strike again, then her credence that he'll strike again will also be 'suspended', as she updates. Again, welcome to the real world.

More directly still, Hájek objects that "when we are deliberating about what we will do, we are *uncertain* about what we will do; probability is our best theory of uncertainty; so offhand, we should assign probabilities to what we will do – or at least, we *may*." He concludes that DARC "is by no means a natural default position; quite the opposite." (511) But Marple remains uncertain whether the butler did it, throughout her afternoon of cogitation. Yet there is a sense in which she is simply 'between' credences on the matter, because she is updating.

2.2. Why terminology matters (I)

An opponent of DARC might object that the Marple analogy doesn't *mandate* the DARC Thesis. There are surely models of Marple's mental processes in which she retains her 75% credence that the butler did it, while she processes the new evidence. Can't we allow that it sits in her mental registers, a record of what she believed before lunch, until overwritten by the results of her stroll in the garden? Can't it simply be flagged, rather than deleted, to mark the fact that she is updating?

This is a welcome point, for it highlights the terminological issue. Trivially, a model of Marple might include records of earlier credences. It is no more surprising that Marple might keep such things than that a flip-tile Arrivals Board might preserve a record of its previous display states.

Having agreed on that, what scope is there for disagreement? The two sides might disagree about whether the states in question really deserve to be called credences. But there is another option, in our view more appealing. The two sides might agree that the dispute about what we call a credence is less interesting than the project of investigating the landscape, to understand why we need to make a terminological choice. Coming at it from the side of those who find DARC attractive, the goal would be to identify features of the epistemic situation of deliberating agents that make it distinctive – that recommend characterisation in terms of DARC *under some understanding of 'credence'*. Once such features have been identified the interesting part of the project will have been completed, in our view, even if there are other factors that recommend modelling deliberating agents so as to allow 'credences' for actions, on other understandings of the term.

We shall come back to these considerations, again exploring them using our updating examples. First we turn to a puzzle raised by Hájek, that in our view provides an clue to what is correct about DARC.

2.3. First-person and third-person perspectives

Two of Hájek's objections to DARC turn on the asymmetry it claims to find between the first and third-person perspectives on the actions of a deliberating agent – or, better, between the 'first-person and present-tensed' perspective and other perspectives. (Proponents of DARC allow that a deliberating agent may assign credences to her actions *at other times*.) The first objection is that the first-person credal gap will ramify, in cases in which the agent is unsure who she is. The second is that the claimed asymmetry gets things backwards – agents typically have *more,* not *less,* information about their own situation, and are therefore better placed than third party observers to assign credences to their own actions.

In both cases, in our view, the analogy with familiar cases of updating does much to defuse the objection. But the second point, especially, will repay further attention. We propose below that DARC be tied to a broader claim about an apparent impoverishment of the first-person perspective on our mental lives.

Hájek puts the first objection like this:

> [C]onsider a ... 'de se' ignorance case, in which you have a lot of evidence regarding an agent's decision situation, but you are uncertain who the agent is. You know a lot about their options ... and so on, and on that basis form credences about what they will choose. Then, you suddenly discover their identity: the agent is *you!* It seems odd that your credences should suddenly vanish. (512)

However, imagine that it suddenly occurs to Marple that the butler is a man she noticed in a hat shop the previous summer, trying on a bowler. That man had not figured in her deliberations; her implicit credence that he was the murderer was close to zero, presumably. Now it suddenly acquires the same status as her credence that the butler did it – i.e., 'suspended and under active consideration', in the model we are considering (though the discovery will presumably lead to a sudden change in any model).

It might be objected that this analogy misses what is most puzzling about Hájek's case. Marple learns something new about the man in the hat shop – that he is the butler – relevant to whether he is the murderer. In Hájek's case, in contrast, "everything you know about yourself had already figured in your reasoning about what 'the agent' will do, it is just that you didn't realize that the information was about yourself."[4]

However, Hájek's own second objection seems helpful here:

> [Y]ou typically have *better* evidence ... than onlookers do regarding your choice: the evidence of your own volition, too. Introspective evidence about your own volition provides *more* basis for the assignment of credences that you might use, beyond publically available information that others may ... use. ... You get to make the news, and you get to know it ahead of everyone else. ... It seems odd that *they* may assign credences

4. Thanks to a referee for the objection and this formulation.

> to what you will choose, but *you* may not, when your evidential basis is superior to theirs. (515)

If there is such a first-person epistemic bonus, it might explain how in Hájek's de se case it simply isn't true that everything you now know about yourself has already figured in your reasoning. A first-person bonus couldn't have figured in your reasoning before you learnt that the agent was you.

To see whether this response does the trick, we'll need a better understanding of the distinctive epistemic character of deliberation – more on this in §4 (see especially fn. 7). For the moment, note that BUG seems helpful with the second objection itself. Marple knows more than PC Plod – she has new evidence, unnoticed by Plod. That's why she's the one updating, and why it is her credence, not Plod's, that is 'suspended' during the period in question. Once we note that deliberation is a species of updating – updating about what one oneself will do – it shouldn't be surprising that an agent's privileged access to her own thought processes means that she lacks credences that other observers can hold. Once again, BUG allows us to say that while DARC does lead to credal gaps, it does so in what ought to seem familiar ways, if we consider what other sorts of updating necessarily involve.

This is good news for DARC as far as it goes, but we might wish it went further. We have identified a covering blanket that protects DARC from concerns about credal gaps. But it covers other things, too, and we have said little about what is distinctive about deliberation as a mode of updating, or how the blanket relates to other considerations in support of DARC. Without that, we don't yet have a full response to the suggestion that the hat shop analogy misses a crucial element of Hájek's de se examples.[5] Moreover, as noted in §2.2 , it is still unclear what is at stake – what is meant by terms such as 'credence', for example.

Accordingly, we now do two things. First, to extract further lessons from the examples in §1.1, we look at factors which determine whether something is a BUG. Second, inspired by Hájek's challenge that DARC's first/third-person asymmetry runs in the wrong direction, we identify such an asymmetry in the direction DARC requires. We argue that it is the key to a valid intepretation of DARC, and an understanding of the terminological choices on which that interpretation depends.

3. Dissecting BUGs

In the examples in §1.1, an informational state of some kind is 'absent' during some temporal interval, as updating takes place. The same applies to any epistemic system that takes its guidance solely from the original system, in real time. In the Arrivals Board case, for example, your belief about the arrival time of a flight is only gappy because you happen to be watching the Board, seeking guidance on that very matter,

5. Similarly, we have no response yet to Hájek's claim that DARC violates the requirement "that credences may be revised only on the basis of evidence" (511). As a referee notes, this seems another disanalogy with the Marple case. See fn. 8 for more.

as it updates. If you had checked it before it updated, gone to the bar, and returned later, you wouldn't have experienced the same suspension of belief.

Even in front of the Arrivals Board your belief needn't remain gappy as the Board updates. Perhaps you have other sources of information to fill the gap. Perhaps it is enough to say, "Siri, when does Big Al's flight arrive?", understanding that Siri will be able to access the same source of information that feeds the Arrivals Board. But the Arrivals Board itself doesn't have this option. It is what it displays, as it were.

True, a split-flap Arrivals Board could be built as one part of a composite system, so that the gap in one place is plugged by an alternative source of information elsewhere – while the split-flaps update, an LED screen shows the latest information (understood as presented from the same source). With this in mind, imagine that a philosopher proposes an analogue of DARC for split-flap Arrivals Boards, the thesis that Updating Annihilates Arrival Information. Others disagree on the grounds just mentioned, i.e., that split-flap Arrivals Boards might be built with back-up displays not subject to the same delays. This isn't an interesting disagreement: one side is talking de facto, the other de jure, or they just mean something different by 'split-flap Arrivals Board'. (In this case both the simple and hybrid varieties are entirely conceivable, and the term could reasonably be used for either.)

But for tennis umpiring the hybrid version would not make sense, *as an umpiring system.* One could certainly consult Siri for her opinion, while waiting for Hawk Eye's judgement. But without the authority vested in Hawk Eye, Siri is simply not part of the umpiring system. An alternative system could give Siri authority in place of Hawk Eye, in some circumstances, but that possibility doesn't change the fact that any umpiring system with a non-instantaneous appeal mechanism is bound to exhibit BUGs. The difference with the previous case stems from the need for a single final authority in an umpiring system, or anything like it. Siri can be a time-saving part of the umpiring system if her judgement preempts and renders redundant that of Hawk Eye, but not if she merely predicts Hawk Eye's judgement, and certainly not if she offers a ruling that will be trumped by Hawk Eye's ruling, in the case that they disagree.

These examples will provide useful analogues to deliberation and DARC. There, too, it will be helpful to see (i) that contingencies concerning an agent's construction may be relevant; (ii) that in consequence, disagreement about DARC may be terminological; and (iii) that there may nevertheless be reasons to prefer one terminological choice to the other – indeed, reasons that turn on issues of authority and finality, as in the umpiring case.

3.1. Two models for random choice

Closer to the case of agency, let's imagine systems whose behaviour depends on an in-built randomiser. Perhaps the system incorporates a coin, tossed at intervals. On Heads the system moves left, on Tails it moves right.

The coin toss gives the system new information. It allows it to update its belief about its own upcoming behaviour – immediately, let's assume. This updating is not

incompatible with the system having credences about the result, *while the tossing is in progress.* If the system knows that the internal coin is fair, it will have a 50% credence for each option. If internal factors can influence the fairness of the coin, the system may be sensitive to those, too, and modify its credence that it will soon move left or right.

Call this creature Partial Dice Head. Its ability to hold these credences about its own upcoming behaviour depends on the fact that it can do two things at the same time: toss its coin, and consider evidence about the result of the present coin toss.

Total Dice Head cannot do this. Like someone who has to close their eyes when sneezing, Total Dice Head has to stop assessing evidence when it wants to operate its on-board randomiser. It doesn't have an internal coin, so it unplugs its entire central processor (CPU), and tosses that instead. (A submodule ensures that the CPU gets plugged back in.) Total Dice Head can't consider evidence about the result of the toss during the tossing process because its CPU is simply *offline* at that point. So Total Dice Head's beliefs about its own future behaviour exhibit BUGs, in a way in which Partial Dice Head's do not.

3.2. DARC as a BUG?

We are now able to sharpen the proposal that the credal gaps that DARC associates with deliberation are nothing more than BUGs. If choosing is somehow *incompatible* (in us) with considering evidence about what we will choose – e.g., because both demand our *attention,* in some non-shareable sense – then we will be in the same position as Total Dice Head. Like tossing one's CPU, the process of choosing will make us *incapable* of simultaneously considering evidence about how we will choose.

Let's call this proposal *Single-Track* DARCness. On the hypothesis that the process of choice is suitably "all or nothing" (Total Dice Head being a crude analogue), it makes DARCness a BUG in much the way that our previous examples were BUGs. In other words, it shows how DARC may be true of us, at least de facto. We may be creatures in which the particular kind of updating that choice involves is incompatible with reflective assessment of the likelihood of our own possible actions.

Accordingly, we want questions like these. Is there some part of the process by which we humans decide how to act that might reasonably be called 'choice', or 'deliberation', that involves a temporary suspension of credence about the actions in question? If so, is this merely a contingent feature of our construction, or does it have some stronger foundation? Are there other things in the vicinity that opponents of DARC might reasonably mean by terms such as 'choice', 'deliberation' and 'suspension of credence', and if so, can apparent disagreements be explained by such terminological variance?

With these questions in mind, let's return to where we were at the end of §2, and take up Hájek's challenge that a deliberating agent should know *more,* not *less,* about her own likely actions than any third party.

4. Transparency and First-Person Blindspots

When we think about others, we have no difficulty distinguishing the question of what they believe from the question of whether their beliefs are true. "She believes that P, but ¬P" is unproblematic. So is "I used to believe that P, but ¬P." But as G E Moore noted, the first-person present-tensed version – "I believe that P, but ¬P" – *is* problematic. Along with "P, but I don't believe that P", it is what we now call Moore's Paradox.

Previous writers (Louise 2009, Price 2012) have noted that Moore's Paradox has some affinity with DARC. Someone who says "I believe (to certain degree) that I'll do A, and I'm making up my mind as we speak" gives something of the same sense of taking back with one hand what they gave us with the other. For our purposes, it will be helpful to turn to Richard Moran's insightful (2001) discussion of issues in this vicinity. Moran himself doesn't mention DARC (or DCOP), but the questions he does discuss seem to be closely connected.

Moran (2001, §2.6) begins with the so-called *transparency* of first-person present-tensed thought, attributing the term to 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 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 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 transparency, but we'll highlight a few points. These will suffice, we hope, to show that the notion is closely relevant to the DARC Thesis.[6]

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. What I think we can see now is that the basis for this

6. For related discussions of transparency and the agent's perspective, we recommend Ismael (2012) and Ahmed (2014); though neither makes the link to DARC that we propose here.

> equivalence hinges on the role of deliberative considerations about one's attitudes. For what the "logical" claim of 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. ... [T]he vehicle of transparency ... lies in the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me. (63)

This involves a distinction between two epistemic stances on one's own mind:

> 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 ineliminable in our cognitive lives, in thought as much as in speech. 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)

(He takes this to be a broadly Kantian point.) And he stresses that the distinction between the two stances does lead, in one sense, to an ineliminable asymmetry between the first-person present-tensed perspective and the third-person perspective (including one's perspective on oneself at other times). Distinctions that we draw easily from the third-person perspective are by default inaccessible from the first-person present-tensed perspective.

4.1. A wider DARCness?

We propose a name for Moran's view about the special character of first-person present-tensed thought. We shall call it the thesis that *Deliberation* (about P) *Annihilates Theoretical Enquiry* (about whether one believes that P) – DATE, for short. In case we seem to be trying to establish a parallel by terminological fiat, we stress that Moran himself 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 imply 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, his answer does not somehow lose its first-person reference because his process of answering it conforms to the transparency requirement. What 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)

So from the first-person present-tensed perspective – and uniquely there – the deliberative path to knowledge 'crowds out' the theoretical path (though the content of the knowledge achieved is precisely the same).

We propose that DARC – or the heart of DARCness, in so far as it has one – is the special, practical case of DATE. In effect, Moran shows us the way to a wider DARCness, a view applicable to factual as well as practical deliberation. Moreover, it is a path that rests on what puzzled Hájek about DARC, namely, that it requires that there is a kind of knowledge that is *inaccessible* to agents themselves (though not to anyone else) about their own states of mind. *Theoretically-grounded* knowledge of one's own mind *is* inaccessible to a deliberating agent, according to DATE, because deliberation crowds it out.[7]

7. Returning to concerns in §2.3, this means that there is what we termed a first-person epistemic bonus: deliberation is a distinctive path to knowledge of oneself. Because deliberation crowds out the theoretical path, there is also a first-person cost. The bonus alone suffices to meet the objection that Hájek's de se example is disanalogous to our hat shop example, while the cost undermines Hájek's suggestion that a new de se belief should leave untouched an agent's previous beliefs about the person she now realises to be herself. So long as she is deliberating, that realisation has the effect of crowding out those previous beliefs (in so far as they concern her present action).

4.2. Why terminology matters (II)

Moran's discussion allows us to identify further possible sources of terminological confusion. As we saw, DATE certainly allows that an agent may know her own mind *by deliberating.* As Moran puts it, transparency amounts to "the requirement that I address myself to the question of my state of mind in a *deliberative* spirit, ... and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me." Accordingly, if we understand DARC in terms of DATE, there is no bar to agents who have credences about what they will do, *achieved by prior (perhaps tentative) deliberation.* On the contrary, this will be simply a case of the kind Moran has in mind, of learning about oneself *by* deliberating – and about this, obviously, the agent herself *is* at an advantage. What is disallowed is reflective, theoretical, evidence-based credence about actions, *in the course of* deliberation about those actions.[8]

Elsewhere (Liu and Price, 2018), we argue that this confusion underlies Jim Joyce's claimed disagreement with Levi about DCOP. Like Skyrms (1990), Joyce (2002, 2007) offers a stepwise model of decision-making in which partially-formed intentions are represented as beliefs (and hence, being partial, as credences). Joyce notes that he follows Velleman (1989) in thinking of intentions as a special class of beliefs, beliefs that are in an important sense self-justifying. Deliberation in Moran's sense is then represented as the phase in the overall process at which an agent updates these beliefs (repeatedly, in a multi-step case). But Joyce agrees with Levi that these credences play no role in this deliberative phase. As he puts it:

> [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. (Joyce, 2002, 79)

We argue that transparency explains why such credences play no role here (and hence provides grounds, as Joyce himself does not, for his claim that such reasoning would be "fallacious"). The upshot is that Joyce and Levi agree in endorsing DCOP (or DARC) *thought of as a special case of DATE.* The apparent disagreement stems from the fact that Joyce thinks of DCOP as the stronger thesis that beliefs about one's own actions can play no role in the extended process of decision-making as a whole.[9] Moran's careful discussion allows us to pull these claims apart.

DATE does not disallow the Siri case, or a model of deliberation in which an agent stores a record of a previous credence (a previous partial intention, say, in Joyce's framework). So it would be beside the point to object to DARC-as-special-case-of-DATE on the grounds that, as Hájek puts it "credences need not be occurrent. You can have a credence without attending to it or being aware of it." (517) Certainly there are

8. This prohibition accounts for the difference between deliberation and the Marple case noted in fn. 5 – DATE explains why deliberation violates the requirement "that credences may be revised only on the basis of evidence," as Hájek (511) puts it.

9. If we call that extended process 'deliberation', we run straight into another source of terminological confusion.

models of deliberation in which this is so – Siri showed us that this is a trivial matter. But this does nothing to undermine DATE, and hence DARC understood as a special case of DATE.

5. DATE and Updating

To clarify the implications of this interpretation of DARC, let's return to earlier concerns. What does DATE add to the observation that DARC is a BUG? What does it tell us about what *kind* of BUG it is?

To focus our thoughts, consider an example. You are staying in an economical B&B. The first morning, the proprietor asks you whether you will have toast with your breakfast, and whether you expect to order toast tomorrow. The first part of the question calls for a decision about whether you'll have toast today. If you say 'yes' and don't eat the toast, the proprietor will be put out, and within her rights to add it to your bill. If you say 'no', you'll have no right to complain about the lack of toast – you didn't order any. But the second part of the question doesn't call for a similar commitment about breakfast tomorrow. Instead, it asks for a declaration about your present inclinations, or expectations, about how you will decide tomorrow. When you repeatedly say 'yes' to this question and yet fail to order toast the following morning, the proprietor has some grounds for complaint. But if she wants to charge you for the toast, she'll need to change the question (and probably the house rules), so that she can ask you to order your toast a day in advance.

After a week of failed predictions, the proprietor has had enough. In an attempt at sarcasm, she asks you whether you will have toast with your breakfast, and whether you expect to order toast *today*. The first question throws you, as usual, into a deliberation about whether to order toast. The second tries to throw you into a different enquiry, a consideration as to whether you are *likely* to order toast today. It tries to put on your plate (so to speak) the kind of question that the transparency of the first person perspective normally requires you to 'see through'. It asks you to treat yourself as an object of enquiry in a context – the one defined by the first question – that, as the transparency thesis points out, normally renders that stance on the matter inaccessible, from your point of view. (The proprietor's point is that you seem incapable of adopting that stance successfully, even when it isn't ruled out.)

What, precisely, is the difficulty in making sense of the second part of the question? Bearing in mind our examples from §1.1 and §3, what is the source of the difficulty, and how deep does it go – is it de facto or de jure, and is it the kind of deficit that a smartphone in your pocket might remedy?

Here are three possible answers about the source of the difficulty:

1. As in Total Dice Head, the cognitive activity required for deciding whether to order toast simply gets in the way of the activity required to answer the second question – you can only think about one thing at once, as it were (especially before breakfast). This would be the Single-Track route to DARCness (and Siri could help).

2. The first part of the question renders the second part *pointless,* in the sense that if you answer the first part then the second part no longer has a non-trivial answer – so there's no point in tackling them sequentially, and little point in tackling them in parallel, even if Siri were to make that were possible.

3. To whatever extent you have an expectation that you will order toast, it is in flux, while you are actually deciding whether to do so. So even if you (or Siri) were to think about the matter, it would be such a moving target as to render the exercise pointless (for a different reason than 2), in normal circumstances.

These are all *empirical* considerations, *practical* factors that might render it impossible or at least pointless to address the second question while addressing the first. DATE offers us something stronger – as Moran puts it, a *categorical* reason for thinking that the first question excludes the second. By Moran's lights what is really odd about the second question is that in the context provided by the first, it amounts simply to *repeating* the first question. This is the effect of transparency, which is itself a product of a categorical difference between the first-person present-tensed perspective and other perspectives on the same subject matter.

How are we to square this conclusion with the fact that we might have Siri in our pocket, *predicting* whether we are likely to order toast, even as our own deliberation stumbles through the pre-breakfast haze to its conclusion? How can it be a categorical matter that the second question doesn't make sense, when it is clearly an empirical matter whether we possess such a digital assistant?

Moran's answer, presumably, will be that Siri isn't part of the agent *in the relevant sense* – not because she sits in his pocket rather than his skull, but because she doesn't share *responsibility* for the results of his deliberations. Perhaps she is *an* agent, if she bears responsibility for her own deliberations. But even if so, she is no more part of *the* agent (in virtue of being in his pocket) than the proprietor of the B&B is part of *the* agent (in virtue of being in the same room).[10]

It would take us too far afield to explore this notion of agency in detail. For our purposes it is enough to note that there is a familiar philosophical notion of agency in the offing here, and that Moran makes a plausible case that if we adopt it, we are committed to DATE (and hence, we argued, to a version of DARC). But it is entirely in keeping with our eagerness to uncover the possible sources of 'merely terminological' disputes about DARC to concede that the meaning of 'agent' itself is one of them. So call the combination of you and your smartphone an 'extended agent', if you wish. Neither DATE nor DARC will apply to *that* kind of agent; but there's nothing here to trouble a proponent of DATE and DARC who does use 'agent' in Moran's sense.

10. Recall our discussion in §3 about the feasibility of adding Siri to a tennis umpiring system. As we saw, either Siri lacks the authority to be part of an umpiring system (even if housed in the same box, in some sense), or she's part of a hybrid umpiring system, subject to the same BUGs. As in the present case, it is authority that makes the difference.

6. Conclusions

We have argued for three main claims. First, practical deliberation is a form of updating. So long as it takes finite time it will produce credal gaps – BUGs – in much the same way as other (finite) forms of updating. This is a broadly empirical constraint, eliminable in principle by speeding things up, but quite enough to allay Hájek's and Rabinowicz's concerns about credal gaps.

Second, there's a deeper, categorical reason why deliberation necessarily produces first-person present-tensed credal gaps, so long as deliberation is conceived in terms of taking responsibility for one's actions and beliefs, and the agent is thought of, more or less, as that which takes such responsibility. Within this framework, DARC is a special case of DATE, which is itself a consequence of the transparency of first-person present-tensed thought about certain of one's own mental states. (Deliberation remains a BUG, but not your common-or-garden empirical BUG.)

Third, we noted that despite transparency, there are ways to interpret terms such as 'credence', 'agent' and 'deliberation' itself so that DARC does not hold, in the terms so interpreted. One trivial way to do so is to keep track of action credences held *before* deliberation, and to continue to call them credences even though they no longer play their normal role (having been sidelined by deliberation). Another is to carry a good phone in your pocket and a relaxed notion of extended agency in your philosophical vocabulary. Do these things by all means, but – a lesson for all sides – be careful that you make them clear to would-be opponents. You may simply be talking past them.

Acknowledgement

This is an electronic version of an article published in *Australasian Journal of Philosophy*. We are grateful to Richard Moran for helpful feedback on an early version of this paper, and to two anonymous referees for many astute comments. We are also grateful to audiences at the 2017 APA Pacific Division meeting, and at Columbia University, Cambridge University, Peking University, and University of Sydney, where earlier versions of this paper were presented. This publication was made possible through the support of a grant from Templeton World Charity Foundation.

References

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

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

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

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. M. (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.

Levi, I. (1997). *The Covenant of Reason: rationality and the commitments of thought*. Cambridge University Press, Cambridge, U.K. ; New York.

Liu, Y. and Price, H. (2018). Ramsey and joyce on deliberation and prediction. *Synthese*.

Louise, J. (2009). I won't do it! self-prediction, moral obligation and moral deliberation. *Philosophical Studies*, 146(3):327–348.

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

Price, H. (2012). Causation, chance, and the rational significance of supernatural evidence. *Philosophical Review*, 121(4):483–538.

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

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press.

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

Velleman, J. D. (1989). Epistemic freedom. *Pacific Philosophical Quarterly*, 70(1):73–97.