Updating as communication

Sarah Moss
ssmoss@umich.edu

On many traditional theories of belief, your belief state is represented by an assignment of credences to propositions, or sets of possible worlds. If you are rational, your credence distribution will be a probability measure. Traditional theories of belief fit with a standard Bayesian theory of rational belief change: on learning a proposition, you must update your belief state by conditionalizing your credence distribution on the proposition you learn. That is, you must update by assigning 0 credence to those worlds incompatible with what you learn, and re-normalizing your credence distribution over the remaining worlds.

Following Quine 1969, Lewis 1979 argues that we should instead represent your belief state by an assignment of credences to sets of *centered worlds*: world-time-individual triples. For instance, if you have .5 credence that it is 3:00pm, your belief state should be represented by a measure that assigns .5 to the set of centered worlds with that time coordinate. Unlike traditional theories of belief, Lewis's theory does not fit with a standard Bayesian theory of rational belief change. For instance, Bayesian conditionalization preserves certainties. If you update by conditionalizing on the set of centered worlds you learn, it follows that if you are ever certain that it is 3:00pm, you must always remain certain that it is 3:00pm. But clearly this is not what rationality requires. If we agree with Lewis about how to represent belief states, we must develop another set of principles governing rational belief change.

In this paper, I develop a procedure for rationally updating credence distributions over sets of centered worlds. I argue that rational updating can be factored into two steps. Roughly speaking, in forming an updated credence distribution, you must first

1. Thanks to audiences at MIT, the 2009 Pacific APA, and BSPC 2009 for helpful questions and criticism. Special thanks to Andy Egan, Adam Elga, Jim Joyce, David Manley, Dilip Ninan, Bob Stalnaker, Eric Swanson, and Mike Titelbaum for extended conversations about earlier drafts of this paper.

use information you recall from your previous self to form a hypothetical credence distribution, and then change this hypothetical distribution to reflect information you have genuinely learned as time has passed. In making this proposal precise, I argue that your recalling information from your previous self resembles a familiar process: agents' gaining information from each other through ordinary communication.

The updating procedure I develop relies on relationships between two kinds of sets of centered worlds: *de se* and *de dicto* propositions. I will define *de dicto* propositions to be boring sets of centered worlds: sets of world-time-individual triples such that if one triple is in the set, so is every other triple which shares its world coordinate. *De se* propositions are sets of centered worlds that are not *de dicto* propositions. *De dicto* propositions are entirely about what the world is like, while *de se* propositions are also about where you are in the world. In §1, I introduce a claim about how *de se* contents of attitudes are related to *de dicto* propositions. In §2, I use this claim to solve a puzzle about imagination. The discussion in §1–2 provides the foundation for a unified theory of communicating and updating beliefs. In §3, I describe how agents communicate *de se* beliefs. In §4, I argue that rational updating begins with a similar process. In §5, I introduce the rest of a complete procedure for rationally updating credences in *de se* propositions. Finally, in §6, I apply my theory of updating to particular cases. The cases elucidate my theory and highlight features of rational updating that any successful updating procedure must recognize.

1 *De se* and *de dicto* contents

In giving a theory of how you should update your *de se* beliefs, it is helpful to understand how the contents of those beliefs are related to various *de dicto* propositions. In what follows, I will argue that the following claim is theoretically and intuitively attractive: given a *de se* proposition, there is a *de dicto* proposition that is equivalent with that *de se* proposition, given what you believe. In more precise terms:

(PROXY) Given a *de se* proposition, there is a *de dicto* proposition such that for any centered world compatible with what you believe, that centered world is in the former proposition just in case it is in the latter.

Some semantic theories of attitude ascriptions help us find *de dicto* propositions equivalent with contents of *de se* attitudes by endorsing the following claim: that speakers use first-person indexicals to self-ascribe attitudes with *de dicto* contents.[2] For ex-

2. MORGAN 1970 and LAKOFF 1972 were among the first to highlight third-personal readings of embedded first-person pronouns. In setting up my examples, I use a case developed by Kaplan in the late 1970's

ample, suppose Kaplan sees himself in a mirror, without realizing that he is seeing himself. Looking at the mirror, Kaplan sees that his pants are on fire, without realizing that his own pants are on fire. In recounting his experience, suppose Kaplan utters the sentence:

(1) I expected that I would be rescued.

Kaplan can truly utter (1), even though he was not aware of being in danger when he looked at the mirror. In this respect, (1) differs from (2):

(2) I expected to be rescued.

Unlike (1), reports such as (2) are true only if the ascribee has a self-directed attitude. Here is one useful explanation of this contrast: reports such as (2) ascribe attitudes with *de se* contents, while reports such as (1) ascribe attitudes with *de dicto* contents.[3] The content of the expectation ascribed by (2) is a set of centered worlds where the center is rescued. But the content of the expectation ascribed by (1) is a set of centered worlds that is characterized not by any property of the center, but by some property of the person Kaplan sees. Since (1) and (2) ascribe expectations with different contents, these ascriptions can have different truth conditions.

This semantic account of the contrast between (1) and (2) fits well with (PROXY). Suppose that (1) ascribes a *de dicto* attitude while (2) ascribes a *de se* attitude. Kaplan believes that the person he sees is not himself, so the content of the attitude that (1) ascribes is not equivalent with the content of the attitude that (2) ascribes, given what he believes. But normally when a speaker utters (1) and (2), the content of the *de dicto* attitude that (1) ascribes will be equivalent with the content of the *de se* attitude that (2) ascribes, given what she believes. These contents will still be distinct propositions. In particular, only one will be a *de se* proposition. But in normal cases, the centered worlds at which the contents differ in truth value will not be among the centered worlds compatible with what the speaker believes.

In other words: even in normal cases, you have a third-personal way of thinking about yourself. On the semantic account just sketched, this way of thinking about yourself gives rise to *de dicto* attitudes that you use first-person indexicals to self-ascribe. The contents of these attitudes are equivalent with the contents of your *de se* attitudes, given what you believe. Or in fewer words: your *de dicto* attitudes are equivalent with your *de se* attitudes, given what you believe. Furthermore, it is natural to think that you retain your normal third-personal attitudes about yourself, even

and familiar from KAPLAN 1989.

3. For recent semantic proposals developing this explanation, see VON FINTEL 2005, PERCUS & SAUERLAND 2003, and STEPHENSON 2009.

if you acquire other attitudes about yourself upon failing to recognize yourself in a mirror. Even in an identity confusion case, your normal third-personal *de dicto* attitudes are equivalent with your *de se* attitudes, given what you believe. So whether or not you are in an identity confusion case, there is a *de dicto* proposition that is equivalent with any given *de se* proposition, given what you believe.

This discussion highlights an important intuition that is ultimately independent of any semantics of attitude ascriptions: you can always think about yourself in the same way you think about any other person, and in the same way other people think about you. Here is another example, one where no ascriptions are uttered: suppose Kaplan has gotten lost in his sea kayak and calls you for directions. He describes the coastline and you look at maps, but you still cannot figure out where he is paddling. There are two possible worlds compatible with your beliefs as you talk to Kaplan: one where Kaplan is in Bellingham, and one where Kaplan is in Seattle. Just the same possible worlds are compatible with what Kaplan believes. Neither you nor Kaplan can figure out which world is actual. So Kaplan is not only ignorant of *de se* facts; he is ignorant of the same *de dicto* facts as you. Suppose Kaplan is in Bellingham and someone just like him is in Seattle, and suppose Kaplan introduces a name for himself as he is talking to you:

(3) Let 'Dr. Demonstrative' name myself.

In one possible world compatible with what you and Kaplan believe, Dr. Demonstrative is in Bellingham and some other guy is in Seattle. In another world compatible with your beliefs, Dr. Demonstrative is in Seattle and some other guy is in Bellingham. Each possible world corresponds to two centered worlds: one centered on Bellingham and one centered on Seattle. Since Kaplan believes that he himself is Dr. Demonstrative, he can rule out exactly two of these four centered worlds. That is why his *de se* belief that he himself is in Bellingham is equivalent with his *de dicto* belief that Dr. Demonstrative is in Bellingham, given what he believes. The same goes for all of his *de se* attitudes.

Considering the kayaking case, we can see that an even stronger moral holds: there is a *de dicto* proposition equivalent with any given *de se* proposition, given merely what you believe with certainty. For instance, Kaplan could always have some shred of doubt about whether he is the man whose pants are on fire, or even about whether he is David Kaplan. Contrast this with your immediate conviction, on uttering (1) and (2) in a normal case, that if one expectation is satisfied then the other will be. Similarly, on uttering (4), you cannot doubt that your expectation is about yourself:

(4) I expect that I will be rescued.

In just this sense, you are always certain about which person is yourself. Given what you believe with certainty, your *de dicto* beliefs about that person will be equivalent with your *de se* beliefs about yourself. Since Kaplan can be certain that he is Dr. Demonstrative, his Dr. Demonstrative beliefs are equivalent with his *de se* beliefs, given what he believes with certainty.

To sum up so far: the *de se* expectation that Kaplan ascribes using (5) is not equivalent with the expectations he ascribes using (6) and (7), given what he believes with certainty:

(5) I expect to be rescued.

(6) I expect that the man whose pants are on fire will be rescued.

(7) I expect that David Kaplan will be rescued.

But other *de dicto* attitudes that Kaplan has about himself are equivalent with his *de se* attitudes, given what he believes with certainty. For instance, the kind of expectation ascribed in (5) is normally equivalent with expectations ascribed using first-person indexicals (as in (4)) and expectations ascribed using names introduced with first-person indexicals (as in (8)):

(8) I expect that Dr. Demonstrative will be rescued.

In particular, (5) and (8) will ascribe equivalent expectations as long as Kaplan remains certain that he himself is Dr. Demonstrative, as he is when he first introduces the name. The same results hold for *de se* attitudes about temporal location. In addition to thinking about yourself from an impersonal perspective, you can think about your temporal location from an atemporal perspective. Just as with your impersonal thoughts about yourself, the resulting *de dicto* attitudes are equivalent with your *de se* attitudes, given what you believe with certainty.

The present approach distinguishes between the *de dicto* belief that Dr. Demonstrative is in Bellingham and the *de dicto* belief that Kaplan is in Bellingham. It is natural to wonder: exactly which worlds are contained in the contents of each of these beliefs? I endorse an indirect answer to this question: your theory of these *de dicto* beliefs should be informed by your theory of other similar pairs of *de dicto* beliefs. For instance, the content of the belief that Hesperus is bright is a *de dicto* proposition, and it is natural to wonder exactly which worlds it contains. The ensuing debate is familiar. If we say the content of the belief is just the set of worlds where Venus is bright, we seem to neglect differences between Hesperus and Phosphorus beliefs. But if we say the content is some descriptively identified set of worlds, we run afoul of familiar anti-descriptivist injunctions.

The resolution of this debate will determine how we characterize the content of the belief that Dr. Demonstrative is in Bellingham. This is a sense in which solutions to Frege's puzzle apply equally well to puzzles of *de se* belief. Several solutions are compatible with the main claim I have introduced. For present purposes, I will remain neutral between them. The main claim is incompatible only with theories according to which we cannot adequately characterize the contents of Hesperus and Phosphorus beliefs unless we say that these contents are *de se* propositions. In that case, nearly all of our beliefs will have *de se* contents, including Dr. Demonstrative beliefs.[4] In that case, nearly all of our beliefs will have *de se* contents, including Dr. Demonstrative beliefs. In order to make my framework compatible with that kind of theory, one must distinguish between deeply and superficially *de se* contents, and read my claims about *de se* contents as claims about deeply *de se* contents.

The claim that (PROXY) holds for a wide variety of agents is attractive, in part because the claim provides for an elegant and unified account of several complicated and disparate phenomena. So far I have argued that (PROXY) fits well with a simple semantics of attitude ascriptions, and that it provides for a natural account of what happens in cases where you think about yourself in the same way other people think about you. In upcoming sections, I put (PROXY) to work in solving a puzzle about imagination, and in saying how agents communicate and update *de se* beliefs. These applications provide further reason to accept (PROXY) itself.

2 Two ways of imagining

Suppose that it is 3:00 and you are teaching class, and while you are teaching, I ask you to imagine that it is 5:00. There are two very different ways you might respond. For instance, you might play along by saying either of the following:

(9) Then I am in my kitchen, starting to make dinner.

(10) Then my watch is wrong, and all of us must be strangely confused to be here so much later than usual.

Once you decide to respond in one of these ways, it is clear how you should go on with what you are imagining. Either you imagine that two hours have passed and your day has proceeded normally, or you imagine that someone has played a practical joke on you and your students. These responses involve very different kinds of imaginary

4. For instance: Chalmers says that the epistemic intension of an indexical is a *de se* proposition. One might argue that 'Hesperus' is an implicitly indexical expression, and conclude that the epistemic intension of 'Hesperus is bright' is a *de se* belief content. See CHALMERS 2002 and CHALMERS 2003 for further discussion.

scenarios. The acceptability of either response raises a puzzle: what distinguishes these two ways of imagining that it is 5:00?

Outside the pretense, you actually believe the *de se* proposition that it is 3:00. Furthermore, you have a *de dicto* belief equivalent with this *de se* belief, given the propositions that you actually believe with certainty. Suppose you introduce '*H*' as a name for the current hour. In addition to believing that it is 3:00, you believe that *H* is 3:00. The *de dicto* content of this belief is central to our solution of the puzzle. The different ways of imagining that it is 5:00 are fundamentally separated by whether what you imagine is consistent with what you actually believe. In both cases, when I ask you to imagine that it is 5:00, you comply by imagining a certain *de se* proposition. In particular, all centered worlds compatible with what you imagine are in the set of centered worlds whose time coordinate is 5:00. But what you imagine in each case is distinguished by whether you also imagine a certain *de dicto* proposition. In the case where you imagine as in (9), you not only imagine the *de se* proposition that it is 5:00, but also the *de dicto* proposition that *H* is 3:00. In the case of (10), this *de dicto* proposition is not part of what you imagine. In other words, there is an extra constraint on the worlds compatible with what you imagine in (9): the *de dicto* proposition that *H* is 3:00 holds in all these worlds.

Our natural responses to (9) and (10) support my characterization of the difference between these ways of imagining. For instance, it is natural to say that when you accept (9), you are imagining that *some time has passed*. If you are imagining that the actual current time has already passed, you may freely imagine that it is 5:00, while imagining that you correctly identified the actual current time as 3:00. In this case, your *de dicto* belief that *H* is 3:00 is true at worlds compatible with what you imagine. By contrast, it is natural to say that when you accept (10), you are imagining that *the actual current time is not what you thought it was*. In this case, your *de dicto* belief that *H* is 3:00 is not true at worlds compatible with what you imagine.

The same puzzle arises for several attitudes besides imagining. For example, there are two natural ways to suppose the *de se* proposition that it is 5:00, corresponding to two indicative conditionals:

(11) If it is 5:00, then I am in my kitchen, starting to make dinner.

(12) If it is 5:00, then my watch is wrong, and all of us must be strangely confused to be here so much later than usual.

Here the puzzle is to say why both of these very different conditionals can be acceptable. Let us agree with Ramsey 1931 that 'if *p*, would *q*' is acceptable to those who accept *q* after "adding *p* hypothetically to their stock of knowledge" (248). Both (11)

and (12) can be acceptable because there are different ways to add the *de se* proposition that it is 5:00 to your stock of knowledge. In particular, as you suppose that it is 5:00, you may or may not continue to accept the *de dicto* proposition that *H* is 3:00. If you retain your *de dicto* belief, you will accept the consequent of (11). If you give up your *de dicto* belief, you will accept the consequent of (12).

So far I have distinguished ways of imagining and supposing centered contents. The distinctions I have drawn are related to the distinction between belief updating and belief revision often cited in literature on *de se* belief change.[5] In order to accept the consequent of (11), you must update on the antecedent as if some time had passed. In order to accept the consequent of (12), you must instead revise your current beliefs. In both updating and revising, you give up some *de se* beliefs. Updating and revising are distinguished by whether you also give up certain *de dicto* beliefs that your old *de se* beliefs were equivalent with. If you retain your *de dicto* beliefs, you are updating. If you give them up, you are revising. I hope to have forestalled the objection that your *de dicto* belief that *H* is 3:00 is trivial, by arguing that whether you retain such *de dicto* beliefs grounds substantive differences in ways of imagining and supposing propositions. I also hope to have forestalled the objection that your belief that *H* is 3:00 is really a *de se* belief, since you imagine the same *de se* contents in (9) and (10), while you imagine the content that *H* is 3:00 only in the former case. To sum up: our puzzle about imagining gives us reason to think that there are non-trivial *de dicto* beliefs equivalent with your *de se* beliefs, given what you believe with certainty. In particular, retaining such beliefs is what unifies several attitudes: imagining as in (9), supposing as in (11), and updating rather than revising. In what follows, I give another reason to accept such *de dicto* beliefs: as I will argue, they play an important role in a simple unified theory of *de se* communication and updating.

3 Learning from other agents

In §4–5, I develop a theory of how agents should maintain and modify their *de se* beliefs as time passes. On this theory, part of updating resembles another instance of the transmission of centered information: interpersonal communication. Communicating agents may exchange beliefs, even though they distribute their credence over entirely disjoint centered propositions, namely sets of worlds with distinct person coordinates. Similarly, an agent may retain beliefs over time, even though at different times, she distributes her credence over sets of worlds with distinct time coordinates.

Lewis says that many belief contents are *de se* propositions. But these *de se* propo-

5. See Katsuno & Mendelzon 1991 for an influential introduction.

sitions cannot always be what is conveyed in communication. For example, suppose Kaplan believes that his own pants are on fire, and when he tells his sister what he believes, she comes to believe just this same centered proposition. Then his sister would come to believe that her own pants were on fire. This is the same centered proposition that Kaplan believes. But obviously, it is not the information that Kaplan should have conveyed to his sister, in telling her what he believed. Instead she should have come to believe some other *de se* propositions, such as the set of centered worlds where the center has a brother whose pants are on fire. The same goes for the transmission of centered information across times. Suppose I express one of my beliefs by saying 'it is Monday' and one day later I remember this belief. Then I should not come to self-ascribe the property of being located on Monday, but the property of being located on Tuesday. These examples illustrate a *prima facie* tension between two intuitive ideas. On the one hand, we may favor a "package delivery" model of communication, on which what I believe is what you come to believe when I communicate my beliefs. On the other hand, Lewis suggests that I believe *de se* propositions. But when I communicate my beliefs, you do not come to believe the same *de se* propositions that I believe. Instead you come to believe other *de se* propositions, ones that I don't believe.

It is not hard to resolve this tension with notions we already have at hand. There is something that Kaplan believes, that he tells his sister, and that his sister comes to believe. It is a *de dicto* content equivalent with the *de se* proposition that his own pants are on fire, given what he believes with certainty. In coming to believe this proposition, Kaplan's sister does not come to believe that her own pants are on fire. Of course, she may acquire several *de se* beliefs of her own. For instance, she may infer that she herself has a brother whose pants are on fire. But the "delivered package" of the Stalnakerian model is a *de dicto* proposition. Just as we can use indexicals to self-ascribe *de dicto* beliefs, we can use indexicals to convey *de dicto* information.

This theory fits with the Lewisian framework, while respecting our intuitions about the identity conditions of contents conveyed in conversation. Stalnaker 2008 worries that the Lewisian framework conflicts with our intuitions about individuating contents:

> Lewis's account distinguishes contents that ought to be identified. If Rudolf Lingens tells you that he is sad, or that he is Rudolf Lingens, and you understand and accept what he says, then it seems that the information you acquire is the same information he imparted. (50-1)

But Lewis can accommodate this intuition, while still taking belief contents to be sets of centered worlds. If Lingens tells you that he is sad, he conveys a *de dicto* proposition equivalent with the content of his *de se* belief that he himself is sad, given what he

believes with certainty. This proposition is something that Lingens believes, that he conveys, and that you come to believe. Our judgment that we should identify what you and Lingens believe reflects the fact that you both believe this *de dicto* proposition. Stalnaker also worries:

> [Lewis] identifies contents that ought to be distinguished. What I believe when I believe that I was born in New Jersey is something about myself, something different from what my fellow New Jersey natives believe about themselves. What I tell the waiter when I tell him that I will have the mushroom soufflé is different from what you tell the waiter if you decide to have the same thing. (50)

But Lewis may respond that when Stalnaker believes that he was born in New Jersey, he believes a *de dicto* proposition equivalent with the content of his *de se* belief that he himself was born there, given what he believes with certainty. His fellow New Jersey native believes a different *de dicto* proposition. Our judgment that we should distinguish what Stalnaker and his fellow New Jersey native believe reflects the fact that they believe different *de dicto* propositions. Similarly, our judgment that we should distinguish what you and Stalnaker tell the waiter reflects the fact that you convey different *de dicto* propositions to the waiter, even if you use the same indexicals when you order.

This discussion suggests a simple theory of the role your *de se* beliefs play in communication. Each *de se* proposition you believe is equivalent with some *de dicto* proposition, given what you believe with certainty. This kind of *de dicto* proposition is something you convey to your audience, and something they come to believe. Furthermore, your audience already has some *de se* beliefs about their relation to you. So they also come to believe some *de se* propositions: the consequences of their standing *de se* beliefs and their acquired *de dicto* information.[6]

Suppose we are standing in a line. I see that I am just behind you, but I have no idea how many people are ahead of you. Suppose you believe a *de se* proposition: that you yourself are fourth in line. This proposition is equivalent with some *de dicto* proposition, given what you believe with certainty. If you say 'I am fourth in line' to me, then this kind of *de dicto* proposition is something that you convey to me, and something that I come to believe. Furthermore, I already have some *de se* beliefs about my relation to you: that I myself am just behind you in line. So I also come to believe a *de se* proposition: that I myself am fifth in line. So when we communicate, I gain *de se* beliefs: not your beliefs, but the consequences of my standing *de se* beliefs and my acquired *de dicto* information.

6. This is a theory of how agents normally communicate. See Egan 2005 for arguments that speakers use epistemic modals to directly convey *de se* propositions.

The theory of communication I defend is incompatible with a certain understanding of why we need to use *de se* propositions to represent mental states. Consider an example from Perry 1977: the amnesiac Lingens is lost in the Stanford library. He reads many books, but nevertheless "still won't know who he is, and where he is, no matter how much knowledge he piles up, until that moment when he is ready to say, '*This* place is aisle five, floor six, of Main Library, Stanford. *I* am Rudolf Lingens'" (710). Some understand Lewis 1979 to argue as follows: Lingens fails to know some information relevant to his location. He knows every relevant *de dicto* proposition. Hence we must use *de se* propositions to characterize his ignorance.

The theory I defend suggests that we should reject this moral of the Lingens case. Lingens is ignorant of *de dicto* propositions relevant to his location, namely those equivalent for him with his *de se* beliefs about his location. This is just the kind of proposition that you would convey to him if you were to resolve his remaining ignorance by saying to him, "you are Rudolf Lingens." Suppose he answers, incredulously, "I am Rudolf Lingens?" The most natural account of this exchange would say that you have used an indexical to communicate a *de dicto* proposition to Lingens, one that you knew and he did not. This account of your exchange is compatible with our pretheoretical description of the Lingens case. Our pretheoretical description may entail that Lingens fails to know some *de se* information. But it does not preclude the claim that Lingens' *de se* ignorance is accompanied by *de dicto* ignorance.

The existence of ignorance does not force us to use *de se* propositions to represent mental states. But the theory I defend is compatible with a second motivation for introducing *de se* propositions, namely that such propositions play a distinctive role in our cognitive economy. Perry 1979 argues that we must mention a certain kind of belief to explain why someone spilling sugar stops pushing his own shopping cart through the supermarket. Suppose that Kaplan and I both believe that Dr. Demonstrative is spilling sugar and both want the spilling to stop. It still remains to be explained why I am motivated to wave and point my finger, while Kaplan is motivated to inspect the cart he is pushing. I have argued that Kaplan has *de dicto* beliefs that are equivalent for him with his *de se* beliefs. But I have not argued that such *de dicto* beliefs obviate reference to *de se* beliefs in explaining action.

4 Learning from your previous self

Giving a theory of how agents with *de se* beliefs communicate illuminates how agents maintain and modify their *de se* beliefs over time. The model of updating I will give relies on an intuitive notion of *genuine learning*. Everyone recognizes that as you sense

that time is passing, you should change your credences to reflect your awareness of your changing temporal location. And your opinions about exactly how much time has passed should influence how you update. But ordinarily as time passes, you are not merely sitting in a black box, keeping track of the minutes as they pass by. You have experiences that make you more informed than your previous self, imposing novel constraints on your credences. In other words, you genuinely learn information. In what follows, I will take for granted the distinction between updating in a black box, and updating as you genuinely learn information.

In black box updating, you form beliefs on the basis of information you get from your previous self. Getting information from your previous self is just like getting information from other agents. Each *de se* proposition you used to believe is equivalent with some *de dicto* proposition, given what you used to believe with certainty. This kind of *de dicto* proposition is something you can currently believe. Furthermore, you currently have some *de se* beliefs about your relation to your previous self. So you can also currently believe some *de se* propositions: the consequences of your current *de se* beliefs and your old *de dicto* information.

Suppose you used to believe a *de se* proposition: that it was the fourth of the month. This proposition is equivalent with some *de dicto* proposition, given what you used to believe with certainty. This kind of *de dicto* proposition is something you can currently believe. Furthermore, you currently have some *de se* beliefs about your relation to your previous self: that your current self is located one day later. So you can also currently believe a *de se* proposition: that it is the fifth of the month. Just as an agent may have certain *de se* beliefs once she acquires *de dicto* beliefs from other agents, you may have certain *de se* beliefs once you recall the *de dicto* beliefs of your previous self.

5 Rational updating: a more complete procedure

Genuine updating happens in two steps. First you update as if you were in a black box. Then you conditionalize your resulting credences on what you genuinely learn. I have sketched how the first step of updating goes. In order to describe genuine rational updating, I will discuss three ways in which the procedure I sketched is idealized, and how these idealizations can be removed.

5.1 Credences

So far I have talked about modifying beliefs, rather than credence distributions. But my aim is to develop a general theory of how agents maintain and modify *credences*.

Fortunately, an appropriately sophisticated theory of interpersonal communication can again serve as our guide. In making an assertion, you can do much more than simply convey certain *de dicto* beliefs to me. If you say 'John smokes' to me, then I should believe that John smokes. But if you merely say 'John *might* smoke' to me, then you merely propose that I should believe that John might smoke. On some recent theories of modals, this means I should give at least some amount of credence to the proposition that John smokes. If you say 'it is .9 likely that John smokes' to me, then I should give .9 credence to the proposition that John smokes. If you say 'if John smokes, it is .9 likely that Mary drinks' to me, then my conditional credence that Mary drinks given that John smokes should be .9. By making assertions, you propose that my credences satisfy some constraint, presumably one that your credences already satisfy.[7]

The analogy with updating extends: in black box updating, your current credences should satisfy constraints that your past credences used to satisfy. Earlier I said that *de dicto* beliefs are what you convey in conversation and recall from your previous self. But in fact what you convey and recall are constraints on your credences in *de dicto* propositions. Suppose you used to give .9 credence to a *de se* proposition: that it was the fourth of the month. Given what you used to believe with certainty, this proposition is equivalent with some *de dicto* proposition, to which you also gave .9 credence if your credences were probabilistically coherent. If you are updating in a black box, you should currently give .9 credence to the same *de dicto* proposition.

Black box updating is like communication: it as if your previous self could talk to you and thereby propose constraints on your *de dicto* credences. Only unlike cases of real communication, there is no limit to the amount of information your previous self can convey. It is as if your previous self proposes that your current *de dicto* credences satisfy every constraint that they did before. So in a hypothetical black box updating case, a case where no genuine learning occurs, all of your *de dicto* credences should stay just the same.

5.2 Conditional credences about your relation to your previous self

So far when talking about how your previous *de dicto* beliefs should influence your current *de se* beliefs, I have talked about your beliefs about your relation to your previous self. But in fact you have more complicated opinions about your relation to your previous self. In particular, your credences about how much time has passed between you and your previous self are conditional in nature. For example, suppose

7. See Swanson 2006 and Yalcin 2007 for developed theories that relate asserted contents to constraints on credences.

you recently looked at a clock that read 2:00, but you think the clock may be an hour early. Suppose you also know that time passes more quickly as the afternoon wears on. Then you might currently believe that if it was indeed 2:00 earlier, four minutes have passed since you looked at the clock. But if it was 3:00, five or six minutes may have passed. In this way, your opinions about how much time has passed are conditional credences. They are conditional on *de dicto* propositions, such as the *de dicto* proposition you would have used 'it is now 2:00' to convey when you were looking at the clock.

In practice, your opinions about your relation to your previous self are given by conditional credence distributions. Earlier your *de dicto* credences were defined on an algebra generated by some partition of atomic *de dicto* propositions. For any such *de dicto* proposition, you have a credence distribution over *de se* propositions, given that *de dicto* proposition. For example, conditional on your having looked at the clock at 3:00, you may give .5 credence to five minutes having passed and .5 credence to six minutes having passed. Conditional credence distributions like these are more precise models of your opinions about your relative location in time.

In black box updating, your credences are entirely determined by two elements: your previous credences in *de dicto* propositions, and your current conditional credences about your relation to your previous self. First your previous credences determine how much credence you give to any given *de dicto* proposition. Then your conditional credences determine how you distribute that credence among all *de se* propositions entailing that *de dicto* proposition.[8] This uniquely determines a credence distribution over both *de dicto* and *de se* propositions. If your previous opinions and your innate sense of time passing were your only sources of information, your rationally updated credences would be determined in just this way.

5.3 Genuine learning

Once we understand how you should update in a black box case, describing a complete procedure for rational updating is straightforward. In ordinary cases, your later credences are not only informed by your previous opinions. They must reflect what you genuinely learn as time passes, information that makes you smarter than your previous self. The combination of your previous *de dicto* credences and conditional

8. Over time, your *de dicto* credences are defined over an increasingly fine-grained algebra, as you come to assign credence to *de dicto* propositions equivalent with the contents of your later *de se* attitudes. So when you distribute credence among *de se* propositions, you are effectively also distributing credence among the *de dicto* propositions that are equivalent with those *de se* propositions, given what your later self believes. See Moss 2011 for further discussion of special problems regarding updating as you expand the range of possibilities over which your credences are defined.

de se credences is a hypothetical credence distribution, representing how you should have updated if you had not genuinely learned anything. In order to arrive at the updated credences you really should have, you must conditionalize this hypothetical credence distribution on what you genuinely learn.[9] It is important to notice that the first step of updating results in a merely hypothetical credence distribution. For example, it may be that you are *always* genuinely learning information, so that you never have credences informed only by your own sense of time passing. Hence rationally updated credences may *always* be the product of your black box credences and what you genuinely learn.

Distinguishing steps of updating that use different kinds of information allows us to more easily recognize how those steps of updating are related to other processes. The first step of updating is analogous to communication. If you have opinions about how you are related to a speaker, she may convey *de dicto* information that constrains your *de se* credences. If you have opinions about how you are related to your previous self, your previous *de dicto* credences may constrain your current *de se* credences in just the same way. The second step of updating is simply conditionalizing on what you learn. In a sense, we have found that conditionalization is the correct procedure for updating *de se* credences. It is just that we must be careful that we are conditionalizing the correct object on what you learn: not your previous credences, but a hypothetical modification of them.

6 Discussion

I have given a framework that organizes and highlights various features of the updating process. The most dramatic consequence of my framework is that the process of rational updating can be entirely factored into two steps: generating hypothetical credences informed only by your previous opinions and your sense of time passing, and conditionalizing these credences on what you genuinely learn. In other words, two kinds of information inform your later credences. There is information you gain from your innate sense of time passing, and there is genuinely learned information that makes you more informed than your previous self. I have argued that these different kinds of information should play different roles in rational updating.

Comparing updating with communication gives us a new way of understanding abnormal updating cases. For example, consider the case of Shangri La from Arntzenius 2003: a fair coin toss determines the way you travel to Shangri La. If the coin

9. This may involve updating by simple conditionalization, Jeffrey conditionalization, or more complicated procedures à la §5 of Diaconis & Zabell 1982.

lands heads, you go on a path by the mountains; if tails, you go on a path by the sea. If you travel on the mountain path, nothing special happens when you arrive at Shangri La. But if you travel on the path by the sea, your memory is erased upon your arrival and replaced by a memory of traveling on the mountain path. Intuitively, even if you travel on the mountain path, you should have .5 credence when you get to Shangri La that the coin landed heads. This is a case of abnormal updating: once you arrive in Shangri La, you can no longer be sure that you traveled on the mountain path, because you can no longer trust your apparent memory.

According to my theory, cases of abnormal updating are best understood by analogy to cases of abnormal communication. Cases where you forget something are like cases where you fail to hear or understand your interlocutor. Cases where you think your memory is anti-reliable are like cases where you think someone is lying to you. Cases where you are not sure whether your memory is anti-reliable are like cases where you are not sure whether someone is lying. Our theory of the latter cases should inform our theory of the former.

Theories of normal communication can be extended to cover cases of lying. For instance, it is natural to think that if you are certain that your interlocutor is lying to you, you should come to believe the negation of whatever he says. If you have some credence that your interlocutor is lying, you should believe a weighted compromise of whatever he says and its negation. Finally, consider the following case: you ask John whether he is a spy. John is either completely corrupt or completely straight: you are certain that John will tell you the truth if he is not a spy and that he will lie if he is a spy. Here you should believe a weighted compromise of his assertion and its negation, where your credence in his assertion is informed by your independent evidence about whether John is a spy. Once you arrive at Shangri La, you are certain that your memory of your journey is reliable if you traveled by the mountain path and that it is anti-reliable if you traveled on the path by the sea. So when you consult your recent memory, it is as if you are communicating with someone who might be a spy: your credence that you traveled on the mountain path must be informed by your independent evidence about which path you took, namely your trustworthy memories from before your journey.

The framework I have given is more modest than some alternative theories. I accept the information you get from your previous self as a primitive input in updating. Since I have not given a general prescription for generating this information, my theory is best understood as a framework within which more detailed updating procedures can be developed. I also accept as primitive the distinction between black box updating, and updating as you genuinely learn information. In other words, I accept

as primitive the distinction between information you gain from your sense of time passing, and genuinely learned information. I take it that we have sufficient intuitive grasp of this distinction for my theory to issue verdicts about particular cases, and I argue for my theory on the grounds that it yields better verdicts about such cases than competing theories. Other theories generally do not distinguish the kinds of constraints on credences that are inputs to updating. For example, Titelbaum 2008 calls the inputs to his updating procedure "extrasystematic constraints," and says only that they "represent rational requirements derived from the specific details of the story being modeled" (560). A number of extant theories similarly do not recognize information you gain from your sense of time passing as a primitive input to an updating procedure. They simply accept as an input that your later self meets certain conditions, such as being certain of the *de se* proposition that it is 5:00, without recording whether you arrived at this certainty by getting additional evidence or by independently keeping track of how much time had passed.[10]

In order to illustrate how my theory works, I will conclude by discussing a very specific case.[11] Continuing the fairy tale motif, suppose the mermaid Ariel has been given the chance to live as a human for three days. On land she loses track of time, so she is unsure whether it is Thursday, Friday, or Saturday. Say she has $1/4$ credence that it is Thursday, $1/4$ credence that it is Friday, and $1/2$ credence that it is Saturday. Suppose that she goes to sleep, and before learning anything else upon waking up the next day, she realizes that it is not yet Sunday. Intuitively, she should then have $1/2$ credence that it is Friday, and $1/2$ credence that it is Saturday.

The framework I have given yields this verdict. Suppose that instead of waking up to learn that it is not yet Sunday, Ariel wakes up in a black box. One day ago, she had $1/4$ credence in the *de dicto* proposition that she would have used 'today is Thursday' to convey, namely that it was Thursday.[12] If she wakes up in a black box, she should still have $1/4$ credence in this proposition (§6.1). Furthermore, conditional on the proposition that it was Thursday, she is currently certain that it is Friday. So she must currently have at least $1/4$ credence that it is Friday (§6.2). Similarly, she must have at least $1/4$ credence that it is Saturday, and $1/2$ credence that it is Sunday. So in a black box case, her credences about what day it is when she wakes up should simply be shifted forward by one day. Ariel should update by conditionalizing these shifted credences on what she genuinely learns when she wakes up: that it is not

10. For other examples, see Kierland & Monton 2005, Halpern 2006, Bostrom 2007, and Meacham 2008.

11. See Arntzenius 2003 and Bradley 2008 for structurally similar examples.

12. Here I adopt the following conventions: 'that it is Thursday' refers to a *de se* proposition, namely the set of centered worlds centered on Thursday, and 'that it was Thursday' refers to the *de dicto* proposition she would have used 'today is Thursday' to convey before waking up.

Sunday (§5.3). Hence my framework confirms our intuition that on waking, Ariel should have $1/2$ credence that it is Friday and $1/2$ credence that it is Saturday.

Other theories have more trouble yielding this verdict. For instance, the Ariel case presents a problem for the updating theory in Titelbaum 2008. I will not examine the problem in detail here, as Titelbaum discusses it at length in Titelbaum 2011 (cf. §8.3.1, "The Sarah Moss Problem"). In short, Titelbaum has trouble with the Ariel case because nothing in his theory distinguishes *de se* propositions about what day it is from *de dicto* propositions about what day it was on a particular occasion. In ordinary cases, the former propositions are conveniently distinguished from the latter propositions, because only the former include propositions in which we should lose certainty as we update. But the Ariel case demonstrates that we cannot rely on this generalization, since Ariel does not lose certainty in any *de se* proposition about what day it is.[13] Nevertheless, what she learns and what she remembers play different roles in updating: her *de se* information that it is not Sunday should supercede her *de dicto* memory that a certain day—namely yesterday—might have been Saturday. In conclusion: a theory of updating must explicitly distinguish your credences about what day it was, what day it is, and how much time has passed. This is what my framework does. Once we recognize that different credences inform your updated credence distribution in different ways, our theory will naturally yield intuitively correct verdicts about cases of rational updating.

13. As a result, both *de se* and *de dicto* propositions are equally qualified to ground our application of Titelbaum's "modeling rule" in the Ariel case. Hence that rule yields inconsistent predictions, namely that Ariel should end up with $1/2$ credence that it is Saturday, and that she should end up with $2/3$ credence that it is Saturday.

References

Arntzenius, Frank. 2003. "Some Problems for Conditionalization and Reflection." *Journal of Philosophy*, vol. 100: 356–70.

Bostrom, Nick. 2007. "Sleeping Beauty and Self-Location: A Hybrid Model." *Synthese*, vol. 157 (1).

Bradley, Darren. 2008. "How Belief Mutation Saves Conditionalization from Self-Locating Information." Ms., Department of Philosophy, University of British Columbia. Available at `http://faculty.arts.ubc.ca/dbradley/`.

Chalmers, David. 2002. "On Sense and Intension." *Philosophical Perspectives*, vol. 16: 135–182.

—. 2003. "The Nature of Narrow Content." *Philosophical Issues*, vol. 13: 46–66.

Diaconis, Persi & Sandy L. Zabell. 1982. "Updating Subjective Probability." *Journal of the American Statistical Association*, vol. 77: 822–830.

Egan, Andy. 2005. "Epistemic Modals, Relativism, and Assertion." *Philosophical Studies*, vol. 133 (1): 1–22.

von Fintel, Kai. 2005. "LSA 311: Lecture 11." Handout from LSA 2005, *Pragmatics in Linguistic Theory*.
Available at `http://semantics-online.org/lsa311/lsa311-ho-9.pdf`.

Halpern, Joseph Y. 2006. "Sleeping Beauty Reconsidered: Conditioning and Reflection in Asynchronous Systems." In *Oxford Studies in Epistemology*, Tamar Szabó Gendler & John Hawthorne, editors, vol. 1, 111–42. Oxford University Press, Oxford.

Kaplan, David. 1989. "Afterthoughts." In *Themes from Kaplan*, Joseph Almog, John Perry & Howard Wettstein, editors, 565–614. Oxford University Press, Oxford.

Katsuno, Hirofumi & Alberto Mendelzon. 1991. "On the Difference Between Updating a Knowledge Base and Revising it." *Proceedings of the 2nd Principles of Knowledge Representation and Reasoning Conference*, 387–394.

Kierland, Brian & Bradley Monton. 2005. "Minimizing Inaccuracy for Self-Locating Beliefs." *Philosophy and Phenomenological Research*, vol. 70 (2): 384.

Lakoff, George. 1972. "Linguistics and Natural Logic." In *Semantics of Natural Language*, Donald Davidson & Gilbert Harman, editors, 545–665. Reidel, Dordrecht.

Lewis, David K. 1979. "Attitudes *De Dicto* and *De Se*." *Philosophical Review*, vol. 88: 513–43.

Meacham, Christopher. 2008. "Sleeping Beauty and the Dynamics of De Se Beliefs." *Philosophical Studies*, vol. 138 (2): 245–69.

Morgan, Jerry. 1970. "On the Criterion of Identity for Noun Phrase Deletion." *Chicago Linguistics Society*, vol. 6.

Moss, Sarah. 2011. "Russell's Principle and Rational Updating." Ms., Department of Philosophy, University of Michigan.

Percus, Orin & Uli Sauerland. 2003. "Pronoun Binding in Dream Reports." In *Proceedings of NELS 33*, M. Kadowaki & S. Kawahara, editors. GLSA, Amherst.

Perry, John. 1977. "Frege on Demonstratives." In *Readings in the Philosophy of Language*, Peter Ludlow, editor, 693–714. MIT Press, Cambridge.

—. 1979. "The Problem of the Essential Indexical." *Noûs*, vol. 13: 3–21.

Quine, W. V. O. 1969. "Propositional Objects." In *Ontological Relativity and Other Essays*, 139–160. Columbia University Press, New York.

Ramsey, F. P. 1931. "General Propositions and Causality." Routledge & Kegan Paul, London.

Stalnaker, Robert C. 2008. *Our Knowledge of the Internal World*. Clarendon, Oxford.

Stephenson, Tamina. 2009. *Towards a Theory of Subjective Meaning*. Ph.D. thesis, Massachusetts Institute of Technology.

Swanson, Eric. 2006. "Interactions with Context." PhD. dissertation, Department of Linguistics and Philosophy, MIT.

Titelbaum, Michael. 2008. "The Relevance of Self-Locating Beliefs." *Philosophical Review*, vol. 117 (4): 555–606.

—. 2011. "Quitting Certainties: A Bayesian Modeling Framework." Ms., Department of Philosophy, University of Wisconsin–Madison.

Yalcin, Seth. 2007. "Epistemic Modals." *Mind*, vol. 116: 983–1026.