Agent and Environment

by

Damien Rochford

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2013

Author .....
Department of Linguistics and Philosophy
August 30, 2013

Certified by .....
Robert Stalnaker
Professor
Thesis Supervisor

Accepted by .....
Roger White
Chair, Department Committee on Graduate Students

Agent and Environment

by

Damien Rochford

Submitted to the Department of Linguistics and Philosophy
on August 30, 2013, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This paper is about how agents learn.

There is a picture of learning that is very influential in epistemology; I call it 'the Classical Picture'. As influential as it is, it is a flawed picture of learning, and epistemology is distorted by it. In this paper, I offer an alternative: the Calibration Picture. It is based on an extended analogy between agents and measuring devices.

Epistemology looks very different from the Calibration point of view. Distinctions that are absolute, given the Classical Picture, are relative, given the Calibration Picture. These include the distinction between enabling and justifying roles of experience, the distinction between *a priori* and *a posteriori* knowledge, and the distinction between irrationality and ignorance.

The beautiful thing about the Calibration Picture is that it gives you a precise way to characterise what is absolute, and a precise way to recover Classical distinctions from that absolute thing, relative to a context. In this way, the Calibration Picture enables you to recover much of the power of the Classical Picture, while offering a new way to understand its significance.

Thesis Supervisor: Robert Stalnaker
Title: Professor

Acknowledgments

Firstly, I must thank my family — my mum Carolyn, my dad Peter, my brother Liam and my sister Chiara. They provide unwavering faith and support in the face of scant evidence that it is well placed. Without that support, there would certainly be no dissertation, and I would certainly be no doctor. I am overcome with gratitude, family!

Similarly, I must thank Alison Mahan, my partner for six years. She was there, as nobody else was, at the very lowest points of my PhD. saga. Though we are no longer partners, I forget none of her kindnesses, and will always be grateful.

Slightly less personally: I want to thank the entire MIT philosophy community, from the years 2006-2013. It's a little embarrassing to say, as it has become a cliché in the philosophy world, but it's true: the MIT philosophy department is a rare and beautiful combination of powerful minds and generous hearts. It has been a gift to be part of such a community. I hope to foster a similar spirit wherever I find myself in the future.

Special thanks go to my thesis committee: Robert Stalnaker, Agustín Rayo and Stephen Yablo. If you know anything about contemporary philosophy, then I am sure you gave a quiet, involuntary gasp of admiration when you read that list of names. I give that quiet gasp daily. That these three minds should all have spent time thinking about something I wrote is more than I could reasonably have hoped for, when I started in this line of work. Again, gratitude.

Finally, I want to thank my thesis supervisor, Bob Stalnaker in particular. It has been an honour being Bob's student. He has the deepest insight of any philosopher I know. But more impressive even than that: his insight is always deployed in a calm, generous spirit of communal inquiry, and never as a weapon. Bob is my model philosopher, and my model human being who does philosophy.

1 The Problem

What is learning? Here is one answer:

> THE CLASSICAL PICTURE (FIRST PASS)
>
> Learning comes in two types. The first type of learning involves acquiring new information through experiences — for example, by looking around and seeing what's going on. Call this *learning from experience,* The second type of learning involves making the most of information you already have — for example, by deducing a conclusion from premises you already know. No looking around required for this second kind of learning. Call this *learning from inference.*
>
> Learning from experience is happening constantly; you are doing it whenever you are having experiences. Learning from inference happens in fits and starts, somewhat haphazardly — at least, it happens haphazardly in human beings, imperfect creatures that we are. So in us imperfect human beings, learning from experience gets mixed up, chronologically, with learning from inference. This means that it can be messy to separate learning from experience and learning from inference, when looking at the learning-history of a real human being. But it is helpful, in understanding the structure of learning, to idealize humans in ways that make the two types of learning more distinct. In particular, we can imagine a certain kind of agent — *the ideally rational agent* — who accomplishes all learning from inference automatically. Such an agent starts off knowing all the things you can know with *no* information provided by experience, and automatically makes the most of any new information she receives.
>
> By considering the ideally rational agent, we can get clear on what kind of learning is from experience, what is learned from inference, and what different norms apply to the different kind of learning.

The Classical Picture seems pretty plausible, right? In fact, the Classical Picture might seem like a slightly worked out statement of some platitudes, rather than a *view* about which we might disagree. Something in the vicinity of the Classical

Picture has been taken for granted for most of the history of philosophy. It was common ground in the classic debate between rationalists and empiricists. Things have gotten more complicated since Quine, who was an influential pioneer in skepticism about the Classical Picture.[1] But the Classical Picture remains highly influential. Bayesian epistemology, on its most straightforward interpretation, is just a particular way of making the Classical Picture more precise, for instance.[2] And many contemporary philosophers explicitly endorse something very close to the Classical Picture. These include Robert Audi, Laurence BonJour, David Chalmers and Christopher Peacocke.[3]

What's more, distinctions whose significance are standardly taken for granted in epistemology depend on the Classical Picture. For example, a standard way of thinking about the distinction between *a priori* and *a posteriori* knowledge presupposes the Classical Picture. Why? Because *a priori* knowledge, on this way of thinking, is just the vacuous case of the second kind of learning — i.e., learning from inference. It's the knowledge you can get by inferring from no experientially-received information at all. *A posteriori* knowledge, on the other hand, is everything else.

Relatedly, the good standing of the distinction between experiences that *justify* knowledge and experiences that merely *enable* knowledge depends on the Classical Picture. An experience merely enables your knowledge if having it enables you to make the most of information you already have; it justifies your knowledge if your knowledge is possible because of information your experience provides you.

Finally, as I alluded above, a standard idea of what *rationality* is presupposes the Classical Picture. An agent is rational in so far as she makes best use of the information she already has, according to this picture. The ideally rational agent

[1]Quine was skeptical about the Classical Picture in a variety of ways, though never explicitly in the terms I have put things. His skepticism about the *a priori/a posteriori* distinction is closely related (see Quine (1953)), as is his view that there is no 'implicit sub-basement of conceptualization' (Quine (1960), p. 3).

[2]This is a fact on which David Chalmers repeatedly relies when defending views that are closely related to the Classical Picture. The best example of this is Chalmers (2011), which is a response to Quine (1953). As this and further footnotes will make clear, Chalmers holds views diametrically opposed to those propounded in this paper; his views form a useful contrast to mine.

[3]See, respectively, the beginning of chapter 8 of Audi (1998), chapter 1 of BonJour (1998), the whole of Chalmers (2012), but especially the fourth excursus, and chapter 1 of Peacocke (2004).

is an agent who wrings everything there is to wring out of the information she's got.

So the Classical Picture is both highly plausible-seeming, at first glance, and highly influential in philosophy. **The problem is**, the Classical Picture is a bad picture of learning. What counts as making the most of information the agent already has and what counts as learning new information from the environment seems clear enough in the cases philosophers typically treat as paradigmatic. But stray a little from the paradigm cases and the distinction becomes pretty dark. Here is an example that brings that out particularly clearly; it is based on a true story.[4]

TEMPERATURE

> Damien grew up in Australia. In Australia, Damien learned how to estimate the ambient temperature from the feel of it. Under conducive conditions, Damien could reliably report the ambient temperature in degrees Celsius fairly accurately — within a couple of degrees Celsius, say.
>
> Damien moved to the United States for graduate school. There he encountered, for the first time, the Fahrenheit scale. After a little while, he got used to hearing and, eventually, producing reports of the ambient temperature in degrees Fahrenheit. After some time, with practice, he could reliably report the ambient temperature in degrees Fahrenheit fairly accurately — within five degrees Fahrenheit, say, under conducive conditions.
>
> Damien was never told anything explicit about the relationship between degrees Fahrenheit and degrees Celsius. One autumn day, he considers that relationship; he realizes he's at an almost total loss. To rectify the situation, he takes stock of the ambient conditions, and estimates the temperature to be in the mid 50s Fahrenheit — about 55° F, say. He then judges the temperature in degrees Celsius to be in

[4]Besides the truth, this example is inspired by examples Timothy Williamson uses to make related points. See Williamson (2007) and Williamson. Stephen Yablo also has examples closely related to this one in Yablo (2002).

> the low teens — about 13° C, if he had to guess. He thus comes to know that when something's temperature is 55 degrees Fahrenheit, it is about 13 degrees Celsius.

Now, when Damien learned that when the temperature is 55° F it's about 13° C, was he making the most of information he already had, or was he learning something new from experience? There's clearly *some* sense in which he was learning something new; before he took stock of the ambient temperature, Damien couldn't tell you nearly as accurately what 55° Fahrenheit is in degrees Celsius, and afterwards he could, and his experience of the ambient temperature is obviously relevant to this change. Damien may well never have experienced ambient temperatures that elicit the judgment of 55° F before; we can stipulate this. So it is not as is if the experience is just awakening some memory. In fact, we can stipulate that Damien had no memories concerning past Fahrenheit or Celsius judgments at all; they got removed accidentally during brain surgery, say. That being so, Damien could have wracked his brain for as long as he liked trying to figure out what 55° Fahrenheit is in degrees Celsius, but without the necessary experience he would have failed. Adding processing power doesn't seem to help. One is tempted to say: the information just isn't there for him to get.

And yet, there is also clearly a sense in which Damien had the information already, even in the case where he lacks relevant memories. For a start, what Damien learned appears to be necessary, maybe even analytic, if there is such a thing, and it is not obvious how experience can teach you about necessary or analytic truths. But put that point to the side; there are other reasons to think that Damien already had the relevant information. Not everyone could learn what Damien did by having the experience he had, including past versions of Damien; what put Damien in a position to learn what he did was that he was already pretty good at both Fahrenheit and Celsius judgments, and it's entirely Damien's past experiences that made this so. And there is a sense in which what Damien learned is independent of the experience he had: his judgment that the ambient temperature is 55° F might be way off. Unbeknownst to him, Damien may be feeling temperatures in a weird way because he's getting sick, or because that terrible surgeon nicked some other part of his brain. But if Damien's temperature-

feeling is distorted, his Fahrenheit and Celsius judgments will be affected by this in the same way, and if they were both very accurate and reliable before Damien's malfeasant surgeon got her hands on him, his judgment about the relationship between Fahrenheit and Celsius will yield knowledge anyway.

In short, it just doesn't seem informative to classify Damien as either learning something new from experience or making use of information he already had; we are missing what's important. TEMPERATURE is an anomaly for the Classical Picture.

If TEMPERATURE were an isolated case, the imperfection of the Classical Picture would not be a big deal. But the isolated cases are those for which the Classical Picture is a comfortable fit, not those for which it isn't. The Classical Picture fits well with certain special cases: learning via explicitly deductive reasoning on one hand, and learning via perception about one's immediate environment or via explicit testimony on the other. But almost everything we know we learn in a way that does not fall into those neat categories. Mostly, we learn things in a messy, in-between way like the way Damien learns that 55° F is about 13° C. Here are some random examples:

> MUSIC: Momoko is musically inclined and plays piano quite well, but has received only the most basic formal training. She can reliably hum the first note of a piece she knows well at the right pitch. But she cannot name the note she hums, and she would fail a written test that required her to name notes. One day, she is thinking about Beethoven's Fifth Symphony. She hums the first note, walks over to her piano (whose keys are marked with the names of their corresponding notes), and hits the keys until she alights on the one that matches the note she is humming. "Ah," she says, "Beethoven's Fifth Symphony starts on a G".
>
> POLICE: Ari has spent a negligible amount of time in the company of on-duty police officers. He doesn't take himself to know much of anything about police culture; if you were to ask him what police culture is like, he wouldn't have anything informative to say. This is despite the fact he is a big fan of cop shows, such as *Law and Order*; Ari (rightly) has little faith in their veracity. One day, Ari catches an

episode of *The Wire*; it 'rings true', as they say. "Ah," he says, "Police culture is like *this* — at least, more like this than *Law and Order*". He's right.

FACTORIZATION: Bob is excellent at mental arithmetic, but otherwise quite unsophisticated when it comes to mathematics. If Bob were ever asked the question "Are 43 and 37 the prime factors of 1591?", he would quickly and easily give the correct answer: "yes". But that isn't what Bob is asked; he is instead asked: "What are the prime factors of 1591?". To answer this question, Bob, needs some props. He gets 1591 pennies and starts trying to arrange them in a rectangle; after a little while doing this, he is able to answer "The prime factors of 1591 are 43 and 37". [5]

MONTY HALL: Jennifer is a statistician. She works for an actuarial firm analysing risk. In most contexts, you would describe her as knowing decision theory like the back of her hand. If you were to present her with the Monty Hall problem[6] as a problem of determining which act maximizes expected utility, given states of certain probabilities and outcomes of certain utilities, she would find the problem trivial. But that is not how the Monty Hall problem is presented to her; it is instead presented in the usual way. Jennifer is adamant that you should be indifferent between switching and staying. "So the conditional probability of getting the prize, given you switch, is the same as the conditional probability of getting the prize, given you stay?" you ask her. She replies "Er... hmmm... this is a question about conditional probability... Bayes' Theorem... whoa! You're right! You should switch! Amazing!"

I could go on. But instead, I'll give you the formula for producing your own examples. Think of a situation in which you would describe someone as *knowing*

[5]This is an elaboration on a case due to Robert Stalnaker; see 'The Problem of Logical Omniscience II' in Stalnaker (1999).

[6]A well known decision-theoretic puzzle. http://en.wikipedia.org/wiki/Monty_Hall_problem

something in one sense but not another. It could be that someone knows the connection between Celsius and Fahrenheit in one sense, to do with how they feel, but not another, to do with calculating one from the other. Or it could be that someone knows what note Beethoven's Fifth starts on for the purposes of humming it but not for the purposes of telling someone else. Or someone knows what police culture is like when he sees a putative example of it, but doesn't know what it is like for the purposes of producing examples. Or many, many other instances. You can always turn such an example into a different, related example in which somebody learns something in the in-between, anomalous way.

Cases of knowing something one way but not another are extremely common. But such cases aren't only common, they are also important. This is why they litter the philosophy literature. One class of such cases that particularly interest philosophers involves an agent knowing an *individual* in one way but not another; these include Hesperus/Phosphorous, Ortcutt the spy, Lingens in the library, Pierre and his puzzling beliefs about London, and more.[7] Another class of examples in the literature involve knowing a *property* in one way but not another; these include water/H_2O, Mary seeing red, knowing that something is a cassini but not an oval, and examples involved in the literature on 'fragmentation', including David Lewis knowing that Nassau Street is parallel to the train tracks in one context but not another, and an outfielder knowing the trajectory of a ball for the purposes of catching it but not for the purposes of saying where it will land.[8] All of these examples can easily be turned into TEMPERATURE-like examples of in-between learning.

The ubiquity and importance of these ambiguous knowledge cases suggests cases of anomalous learning are equally ubiquitous and important, as whenever an agent can be described as knowing something one way but not another she is liable to learn in the anomalous way. This is a problem for the Classical Picture.

One way we could react to this problem is by trying to reinterpret the anomalous cases. I'm sure a Classical philosopher could come up with some way of

[7]For sources of these cases, see, respectively, Frege (1948), ,Quine (1956), Perry (1977) (which was alluding to Frege (1956)) and Kripke (1979).

[8]Sources are Kripke (1980), Jackson (1982), Yablo (2002)., Lewis (1982) and Stalnaker (1999), respectively.

making Classical distinctions among all the anomalous cases, with a little fancy foot work.[9]. But enforcing Classical distinctions among such cases seems arbitrary and distorting; the cases obviously are more alike than not, epistemically speaking. We wouldn't tolerate the arbitrariness or bother with the fancy footwork if we had a clear alternative to the Classical Picture; it would be like preferring an epicycle-upon-epicycle version of the geocentric theory of the solar system to the heliocentric theory. But what alternative is there?

There have been philosophers expressing scepticism about the Classical Picture in one way or another at least since Quine, but they have not had a lot in the way of positive alternative on offer; hence the persistence of the Classical Picture. Quine himself rejected any pretension to epistemological theorising that wasn't straight-up descriptive psychology.[10] We have made too much progress within the Classical paradigm to be content with that attitude. Just throwing out the entire Classical project is not a viable alternative.

I hope to provide something that *is* a viable alternative in the rest of this paper. One of the reasons it is viable is that the Classical Picture is recoverable as a special case. We can have a proper understanding of the anomalous cases *and* do the work the Classicists want to do, with very similar tools. But, as you will see, we will end up with a very different understanding of those tools.

2 The Solution

The nice thing about TEMPERATURE is that it suggest a certain analogy: an analogy between agents and measuring devices. This analogy is the basis of a way of thinking about agents, their minds, and their intentional and epistemological properties that has got its fullest development in the seminal work of Fred Dretske.[11] This way of thinking about agents is attractive quite independently of the considerations of this paper; I aim to show that it also delivers an alternative to the Classical Picture of learning.

Consider the following three cases of using a thermometer to learn something:

[9]Or, if the Classical philosopher is Chalmers, an unyielding march. See the eighth excursus of Chalmers (2012)

[10]See Quine (1969).

[11]See, in particular, Dretske (1981), and also Dretske (1988).

Case 1: Your thermometer reads in degrees Celsius only. You use it to find out the temperature outside. You do this by taking it outside a while. It reads '13° C'.

Case 2: Your thermometer reads in both degrees Celsius and Fahrenheit. It's an analogue thermometer, and the markings indicating degrees Celsius and Fahrenheit are written on the tube. You use the thermometer to find out that 55° F is about 13° C; you do this by looking at the markings on the tube.

Case 3: You have two thermometers; one reads in Celsius, the other in Fahrenheit. You use the pair of thermometers to find out that 55° F is about 13°; you do this by taking them outside a while, taking a reading on the one thermometer, and then taking a reading on the other.

There seems to be some good sense in which the information that it is 13° C is new information your thermometer acquired from the environment, in Case 1. And there seems to be some good sense in which the information that 55° F is about 13° C is already in your thermometer, before you do anything with it, in Case 2. But what about Case 3? Is the information you learn in this situation something that was in the pair of thermometers already, and the environment merely enabled you to access it, or is it something extra, which involved acquiring new information from the environment? That seems like a silly question. What is true is that the thermometers, calibrated as they were, and the environment being as it was, together made the relevant information accessible; nothing is gained by trying to partition the information available into that which the thermometers together 'already had' and that which required extra input from the environment.

The analogy between the two thermometers and Damien in TEMPERATURE is obvious. Damien is like the thermometers, and the information available to him is not well partitioned into the Classical categories for the same reasons the information available via the thermometers isn't well partitioned into Classical categories.

The analogy between Damien and the thermometers is central to my alternative picture of learning. I call my alternative 'The Calibration Picture'. Though

the analogy between Damien and the thermometers is obviously quite particular, I follow Dretske in thinking that it can be made completely general and that the Calibration Picture is a completely general picture of how agents learn.

To make the contrast with the Calibration Picture clear, it will be helpful to put the Classical Picture slightly more precisely:

> THE CLASSICAL PICTURE (SECOND PASS)
>
> There are two kinds of information an agent can learn at a time t:
>
> (1) there's the information that's available to learn because of the state of the agent's environment at t, and
>
> (2) there's information available to learn because of the internal state of the agent, just prior to t.
>
> An agent learns something at t either by getting new information from her environment at t, or by using the information available to her because of her internal state just prior to t. If she does the first, she learns via experience; if she does the second, she learns from inference.

Here is a rough first-pass at the alternative, Calibration picture; it's a little sparse on details, but I hope it gives you an idea where we are headed.

> THE CALIBRATION PICTURE (FIRST-PASS)
>
> The information available to an agent at a time t is a matter of two things: her internal state at t and the state of her environment at t. But the information available to an agent cannot, in general, be factored into the part which is available because of her internal state just prior to t and the part available because of her environment at t.
>
> What there is instead is a spectrum of cases. Towards one extreme, the agent is internally configured in such a way that certain information is available in a very wide variety of environments; that is a bit like the relevant information being encoded by the agent's internal state. Towards the other extreme, the agent is internally configured in such

> a way that whether certain information is available or not depends very delicately on what her environment is like; that is a bit like the information being encoded in her environment. But most of our epistemic lives is in the middle of the spectrum, far from the extremes.

Here's an analogy to help make the contrast clear. Consider the various characteristics a particular organism can have. In the case of a human being these will include her eye-colour, her lung-capacity, her intelligence, her level of income, whatever. You might have a naïve view of the causes of these characteristics. On the naïve view, the causal contribution can be factored into two parts: the part which is caused by the organisms genes, and the part which is caused by her environment. People often misinterpret claims like 'the heritability of intelligence is 80%' in a way that tacitly presupposes something like this naïve view. The idea is that there is part an individual's intelligence — 80% of it — that is caused just by her genes, and the rest depends on her environment.

The alternative to the naïve view is this: an organism's characteristics are caused by both its genes and its environment, but there is no general way to factor the causal contribution into the part due to the genes alone and the extra part due to the environment. What there is instead is a spectrum of cases. Towards the one extreme, the organism's genes are such that it will develop a certain characteristic in a very wide variety of environments; that's a bit like the characteristic being caused by the organism's genes alone. Towards the other extreme, the organism's genes are such that whether it develops a certain characteristic or not is very sensitive to exactly what environment it ends up in; that is a bit like the characteristic being caused by the organism's environment. But most cases are in the middle of the spectrum, far from the extremes.

The Classical Picture is the analogue of the naïve view, and the Calibration Picture is the analogue of the alternative. The alternative is the correct view, when it comes to the characteristics of organisms. I say that its analogue, the Calibration Picture, is the correct view when it comes to learning.

Here is the plan for the rest of this paper.

It will take some work before the Calibration Picture can be painted with some clarity and rigour. That is the goal of section 3.

In sections 3.1 and B.2, I discuss the calibration of simple, paradigmatic measuring devices, paying particular attention to our pair of thermometers. I will introduce a way of modelling the state of calibration of such devices, and draw lessons that apply to agents also.

In section 3.3, I discuss the different ways in which the environment plays a role when you learn something using a measuring device, using our three cases above as illustration. The model developed earlier enables us to be relatively precise about this. Again, I draw lessons that carry over to agents too.

In section 3.4, I give you a more precise statement of the Calibration Picture; it is made possible by and motivated by the considerations in the preceding sections.

Section 3.5 is a segue from section 3, which focuses on measuring devices, with an eye to lessons for agents, to section 4, with a focus more directly on agents. In section 3.5 I discuss some of the ways in which agents are more complicated than measuring devices, and briefly outline how to extend the formal model of calibration to cover agents (the details are in an appendix).

At this point, we have the picture of learning in place. To do some epistemology, we need to know how *rationality* fits into the picture. That is the topic of section 4.

Because rationality is primarily a matter of an agent's total belief-state, we need to know how to characterise an agent's total belief-state, given her state of calibration, in order to know how rationality fits into the Calibration Picture. That is the topic of section 4.1. This is a less trivial matter than you might expect.

Section 4.1 delivers us a very general way of characterising an agent's total belief-state. After discussing certain ostensible limitations of this way of characterizing an agent's total belief-state in section 4.2, I argue that there is no total belief-state, so characterised, that is distinctive of ideally rational agents; I do this in section 4.3. For any one of these total belief-states, you can come up with a situation in which it rational for an agent to have that total belief-state.

This has far reaching consequences. In particular, it means that there is no context-independent way to draw the line between irrationality and ignorance. In section 4.4 I propose we replace context-independent talk of rationality with a new relativistic notion. My view goes further than just claiming that rationality and knowledge are context sensitive; it gives a characterization of what is invariant

across contexts, and shows how the invariant thing can be decomposed in to a characterization of the agent's state of ignorance and state of rationality relative to context.

Section 4.5 is an application of the proceeding work to a particular argument — sometimes called 'the Frontloading Argument' — which appears to show that we are in a position to have a lot of *a priori* knowledge. If you buy the Calibration Picture, I claim, you will find this argument ambiguous between several readings. On only one such reading will you find the argument sound, but on that reading the conclusion is much less exciting than the fan of the *a priori* might have hoped.

I conclude, very briefly, in section 5.

There are some appendices; they contain some more technical details on the models mentioned in the body of the paper.

3 Calibration of Measuring Devices, and Agents

3.1 Accuracy, Reliability

Consider our pair of thermometers. The fact that we can learn something from them has to do with their state of *calibration.* But what does that mean?

Let's start by thinking about what a thermometer is. A thermometer is something you use to get information about something else. Call this something else 'the system'. The system is what the thermometer measures. It might be a human, or a cooking turkey, or the air in my apartment. A thermometer can be in many different states; for instance, the height of the column of mercury can be one centimetre, or two, and so on. The system can also be in very many states; the relevant ones are the different temperatures the system can be. Each state of the thermometer is supposed to be a good indicator of a particular temperature-state of the system.

What does it mean for a state of the thermometer to be a 'good indicator' of a particular temperature state? It means two things: firstly, that the thermometer state is *accurate,* and secondly, that it is *reliable.* First, consider accuracy. Take the state of the thermometer that is supposed to indicate that the system is at 30° Celsius; call it d. State d is accurate iff the system is close to 30° C whenever the

thermometer is in state d.

Actually, we should relativize accuracy to background conditions, as a measuring device's performance varies a lot with background conditions. Early thermometers only worked well within certain very limited pressure ranges, for instance. The thing to say, then, is that state d is accurate, *relative to certain background conditions*, iff, *assuming those background conditions obtain*, the system is close to 30° C whenever the thermometer is in state d.

The second way in which a thermometer state can be a good indicator has to do with these background conditions. State d is reliable iff the range of background conditions under which it is accurate is large. The reliability of thermometer states was greatly improved by sealing the relevant thermometric material in a glass tube, thus isolating the material from the environmental pressure.[12]

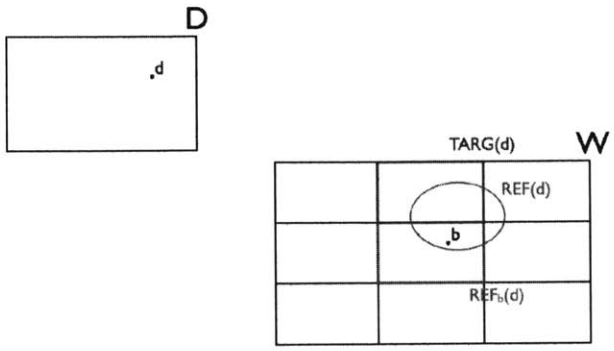
A thermometer as a whole is accurate, relative to certain background conditions, iff its states are, on the whole, accurate relative to those background conditions; it is reliable as a whole iff its states are, on the whole, reliable.

As it is with thermometers, so it is with measuring devices generally. A measuring device is, in general, a thing that can be accurate and reliable, or fail to be; this requires that it have a variety of states that are supposed to indicate things about the state of some system. How accurate and reliable a measuring device is is its *state of calibration.*

We can model a measuring device's state of calibration formally using a set-theoretic structure. In the first appendix I give the official, relatively succinct, algebraic description of such a model. I will basically recapitulate that information here, but in a more geometric way that hopefully makes the model intuitive.

Here is a picture of a model of the calibration of a particular state of a measuring device:

[12] An innovation due to Fernando II de Medici, according to Wikipedia.



The rectangle on the left side, labelled 'D', represents the set of the measuring device's states. Each point represents a particular state of the device; one of them has been highlighted, and labelled 'd'. D could be the set of thermometer states, and d could be the state in which the mercury column is three centimeters high. The big rectangle on the right, labelled 'W', represents the states of the rest of the world, including the system the device is supposed to measure.

W is broken up into cells by a grid; each cell represent a state of the system. The system remains in the same state as you move inside a cell; the only thing that varies is the state of the rest of the world. So, at every point inside one cell, the ambient temperature of my house might be 15° C, but other things will be different — someone walks down the street at one point, not at another, and so on.

The states the cells represent are assumed to have a distance-relation defined on them (what mathematicians call a *metric*). The state the system is in when its temperature is 15° C is relatively close to the state it is when it is at 18° C, and relatively far away from the state it is in when it is at 100° C, for instance.

The blue cell, labelled 'TARG(d)', is the state of the system that d is supposed to indicate — that the temperature of the system is 30° C, say. In general, 'TARG' denotes the function that takes a device state to the system state it is supposed to

indicate, which I sometimes call its *target*. So the state of the system when it is at 30° C is the target of the state of the device when the mercury column is 3cm high, for instance. You can think of d's target as what d represents.

The red oval labelled '$\mathsf{REF}(d)$' represents the set of all world-states that are compatible with the device being in state d. In general, 'REF' denotes a function that takes a device state to the set of all world states that are compatible with it, given its state of calibration. I will sometimes call $\mathsf{REF}(d)$ d's *reflection*.

In the diagram, $\mathsf{REF}(d)$ mostly overlaps $\mathsf{TARG}(d)$, which means that most of the world-states in which the thermometer is 3cm high are world-states in which the system is at 30° C, like it is supposed to be. A way of thinking about how good d's state of calibration is is as how close $\mathsf{REF}(d)$ is to being inside $\mathsf{TARG}(d)$ — or, to put it another way, how close d's reflection is to being inside d's target. Perfect calibration occurs when the reflection is entirely inside the target — i.e., when *all* world-states compatible with the device being in d are ones in which the system is in $\mathsf{TARG}(d)$. For example, our thermometer state would be perfectly calibrated if the only ways for the world to be that were compatible with it were world-states in which my room is 30° C. But in our diagram, d is not perfectly calibrated; there are some world-states in which the thermometer's mercury column is 3cm high, but my room is not at 30° C.

One of the states of the world in W has been highlighted; it is labelled 'b'. That represents the state of the world when particular background conditions obtain and the device is in state d. When that is so, you can see from the diagram that the system is in the state represented by the red cell, labelled '$\mathsf{REF}_b(d)$'. The accuracy of d, relative to background conditions b, depends on how far away $\mathsf{REF}_b(d)$ is from $\mathsf{TARG}(d)$. If we take the physical distances between cells in the diagram as a guide, then the accuracy of d relative to b is pretty good, though not perfect, as $\mathsf{REF}_b(d)$ is close to, but not identical with, $\mathsf{TARG}(d)$.

Moral of this section: a thermometer, and more generally any measuring device, is a thing whose states are supposed to indicate how things are with some system that it measures; it will be in some state of calibration with the system, which we can model. Agents are also things whose states — viz., belief-states — are supposed to indicate how things are with a certain system: the entire world. Agents will also be in some state of calibration with the system, and can

be modeled accordingly.

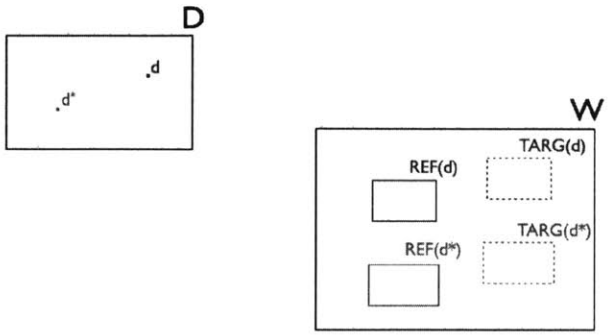
3.2 Relative Calibration

The states of measuring devices stand in what I call *relative calibration* relations with each other in virtue of their calibration with the system. Here's a comparatively simple example. Suppose we have two Celsius-reading thermometers; suppose further they are both well calibrated. Then one will read '30° C' if and only if the other does too. That relationship, of one device-state obtaining if and only if the other does, is a relative calibration relationship.

That's an example of a relative calibration relationship that is as it should be; the two states are *well* calibrated with respect to each other. But relative calibration relationships can fail to be as they should be. It might be that one thermometer reads '30° C' if and only if the other reads '35° C', or it might be that when the one reads '30° C', the other can read anything between '25° C' and '35° C'. These are different, sub-optimal relative calibration relationships.

Relative calibration relationships don't just hold between states of different devices; they hold among the states of a single device too. If everything is going right, then a thermometer will read '30° C' only in conditions that are relatively far away from conditions in which it will read '0° C' — relative, that is, to those in which it will read '25° C' (for instance). If the distance between the conditions in which the thermometer reads '30° C' and the conditions in which the thermometer reads '0° C' are as they should be, then the relative calibration relation between the '0° C'-state and the '30° C'-state is good. But this relative calibration relation could be bad too. For example, the conditions that elicit the '0° C'-state could be too close or too far away from the conditions that elicit the '30° C'-state.

We can use our model of calibration to give a formal definition of the relative calibration between two states. As before, the official statement is in an appendix, and I'll give a more intuitive, geometric understanding of that definition here. Here is our picture:



The rectangle at left, labelled 'D', represents the sets of states of a device. The rectangle on the right represents the states of the world, including the system, just as before. Two of the device-states are highlighted; one is labelled 'd', and is blue; the other is labelled 'd*' and is red. Over among the world-states are the targets and reflections of d and d^*. The targets are labelled with 'TARG's, and outlined with a dotted-line; the reflections are labelled with REFs and are outlined in whole lines. The target and reflection of d are blue; the target and reflection of d^* are red.

Recall that there is assumed to be a metric on the system states; if the relevant states are temperatures, then, according to the metric, the state of being 30° C is close to the state of being 31° C and far from being 100° C. We will take the physical distance in the picture above to represent distance according to the metric; so the physical positions of the blue and red rectangles above represent how distant the relevant targets and reflections are, as measured by the metric.

You can see from the picture how distant the target states of d and d^* are supposed to be. That tells you what constitutes good relative calibration between d and d^*. Roughly speaking, you can see what the *actual* state of relative calibration between d and d^* is by looking at the relative positions of their reflections. If the relative position of the reflections is close to the relative position of the targets,

then d and $d^\star$ have good relative calibration (again, roughly speaking). In the particular case illustrated, you'll see that the relative positions of the reflections is, in fact, the same as the relative position of the targets. This means the relative calibration between d and $d^\star$ is very good. The two states have very good relative calibration, in the case illustrated, even though their straight-up calibration with the system is quite bad, as you can deduce from the fact that d's target is quite far away from its reflection, and the same is true of $d^\star$.

This is an important, general point. If two states are well calibrated with the system, their relative calibration with each other will be good — that follows just from the definition of relative calibration. But the converse is not always true; it is possible for both states to be poorly calibrated with the system, but their relative calibration to be good. It might be that there is a systematic error in the way the one state tracks the system, but that systematic error is replicated in the other state. This is what would happen if the wrong place were marked '0' on an otherwise good thermometer, for instance.

Moral of this section: The states of a measuring device stand in relative calibration relations with each other. These relationships can be good or not so good in degree. If the calibration of the relevant states with the system is good, their relative calibration with each other will be good, but the converse is not always true: the relative calibration of two states with each other can be good even when their calibration with the system is bad.

As it is with measuring devices, so it is with agents. Belief-states stand in relative calibration relations with each other, and these relationships can be good or not so good in degree. If the beliefs are well calibrated with the world, then they will be well calibrated with each other, but the converse is not always so. This is just another, slightly more precise way of putting a very familiar point: that if your beliefs are accurate, they will also fit together well — i.e., they will be *coherent*. But your beliefs can be coherent without being accurate.

3.3 Three Cases and Three Roles for the Environment

Recall the following three cases of using a thermometer to learn:

Case 1: Your thermometer reads in degrees Celsius only. You use it to find out the

temperature outside. You do this by taking it outside a while. It reads '13° C'.

Case 2: Your thermometer reads in both degrees Celsius and Fahrenheit. It's an analogue thermometer, and the markings indicating degrees Celsius and Fahrenheit are written on the tube. You use the thermometer to find out that 55° F is about 13° C; you do this by looking at the markings on the tube.

Case 3: You have two thermometers; one reads in Celsius, the other in Fahrenheit. You use the pair of thermometers to find out that 55° F is about 13°; you do this by taking them outside a while, taking a reading on the one thermometer, and then taking a reading on the other.

As I said in section 2, it looks like there is some good sense in which the information is in the environment in Case 1, in the thermometer in Case 2, and that that way of classifying things seems unhelpful in Case 3.

What accounts for these differences? Let us first focus on Cases 1 and 2, which offer the clearest contrast, and see what lessons we can draw for Case 3. There are, I say, three roles the environment plays in making your learning possible in Case 1 that it does not in Case 2, and these three roles of the environment explain the difference in where we are inclined to locate the information learned in the two cases.

The first, most obvious role the environment plays in Case 1 but not Case 2 has to do with what it is you learn in each case. What you learn in Case 1 is contingent, and what you learn in Case 2 is necessary. To learn something, that something needs to be true in the environment in which you learn it. This puts a substantive constraint on what the environment must have been like, for you to learn what you did in Case 1 — viz., the environment must have been one in which the temperature was indeed 13° C. Similarly, to learn what you did in Case 2 your environment must have been one in which 55° F is about 13°. But that is not a substantive constraint, as *every* environment is one in which 55° F is about 13°.

The second role the environment plays in Case 1 but not in Case 2 is this: unless your thermometer is perfectly reliable (which it isn't, in any realistic case),

the accuracy of the thermometer state that reads '13° C' is sensitive to environmental conditions. The environment needs to turn out right, in Case 1, for the accuracy of the state to be good enough to allow learning. But now consider Case 2. We can suppose the thermometer has a state that reads '55° F is about 13° C' in Case 2 — it is, presumably, a state that obtains all the time. This state will count as perfectly calibrated trivially, as its target includes *every* state of the world (being necessary). So, in Case 2, there is again no substantive constraint on the environment imposed by the requirement that the environment be one in which calibration is accurate, unlike Case 1.

Finally, we come to the third role the environment plays in Case 1 but not case 2; it has to do with relative calibration. Having a perfectly accurate '13° C' thermometer-state is not sufficient to learn that it is 13° C, when it is. It is compatible with this state being *perfectly* accurate that the thermometer nevertheless read something incompatible with it being 13° C — it might read '25° C', for instance. This possibility is avoided only if the relative calibration between the thermometer's '13° C' state and its '25° C' state is accurate enough, in the prevailing conditions. More generally, using the thermometer to learn that it is 13° C requires that the 13° C' state have, in general, accurate enough relative calibration relations in the prevailing conditions. This puts a third substantive constraint on the environment in Case 1.

That constraint applies in Case 2 also. It is easy to miss because it is so easy to tell that the constraint is met — you can just look at the thermometer and see that it is not going to read anything incompatible with the fact that 55° F is about 13° C. So the requirement that the relevant state's relative calibration relations be accurate enough is likely to put substantive constraints on the environment in Case 1, and not in Case 2. This is the third way in which the environment plays a role in Case 1 it does not in Case 2.

These three ways in which learning is much more sensitive to co-operation from the environment in Case 1 than Case 2 explain why Case 1 is well thought of as using a thermometer to extract information from the environment, and Case 2 is well thought of as learning information encoded in the thermometer itself.

Now consider Case 3 — the case involving two thermometers. Case 3 is like Case 2, in that what is learned is necessary, and in that the calibration of the

relevant state is perfect. But it is like Case 1, in that poor relative calibration is a live possibility. The Celsius readings could be more or less well calibrated with the Fahrenheit readings, so the thermometers could well give readings inconsistent with the '55° F is about 13° C' reading. So learning what you do in Case 3 requires the co-operation of the environment, but not in the way we usually think of — not because what you learn could have been false, or because the calibration of the relevant device state is sensitive to the environment. Non-coincidentally, Case 3 does not fit very easily into the information-in-the-device/information-in-the-environment dichotomy.

These observations suggest a certain picture that I will make explicit in the next section.

Moral of this section: There are three ways in which learning something by using a measuring device requires the co-operation of the environment. Firstly, what is learned must be true in that environment. Secondly, the calibration of the relevant device state must be accurate enough in that environment. Thirdly, the relative calibration of the relevant device state with other device states must be accurate enough, in that environment.

As it is with measuring devices, so it is with agents. The three ways that learning via a device requires co-operation from the environment are ways in which learning generally requires co-operation from the environment: what is learned must be true, the agent's belief-state must be well calibrated in the environment, and the relative calibration relations among her belief-states must be good enough in the environment.

3.4 The Calibration Picture

The foregoing observations suggest a picture of learning that applies equally to measuring devices and to agents. When applied to agents, it is a way of making the first pass at the Calibration Picture, in section 2, more precise.

I give the official statement of the view below. Before I do, a point of clarification. You'll note that I talk of 'the information that s', and talk separately about a state of the world p. The 's' here is supposed to be a substitutional variable, where one sticks in a sentence, whereas the p is an objectual variable that ranges

over states of the world[13] — in particular, it refers to the state that makes the corresponding s true. Why this baroque formulation of things, rather than just talking about the information that p obtains?

This complication is necessary to allow for the possibility of individuating information very finely. For example, we want it to be possible that the information that it is 13°C be available to an agent, but not the information that it is 55° F, despite the fact that the very same state of the world makes both 'it is 13°C' and 'it is 55° F' true. That will be possible, on the below formulation of things, because we will individuate information, and beliefs carrying that information, by the sentences used to express that information.[14] This allows for the possibility that the relative-calibration requirement (clause 4 below) for the 'it is 13°C' belief is different from the requirement for the 'it is 55° F' belief.[15]

As a reminder that beliefs are being individuated by the sentences that express the information they carry, I will sub-script the belief variable b below with the sentence variable s.

Ok; here is the official statement of the view.

THE CALIBRATION PICTURE (SECOND PASS)

The information that s is available for an agent A to learn in a context c only if:

1. there is a particular state of the world p such that,

[13]Caveat: it ranges over states of the world if one understands 'state' in a very general way. In particular, it should include any way the whole of space-time can be, rather than ways for the world to be at a particular time only. Following Robert Stalnaker (see Stalnaker (1976), for instance), I am inclined to call such things 'propositions'; hence the use of 'p'. However, the history of philosophical usage of that word makes it muddied with complicated and possibly misleading connotations; I think it is less misleading if I talk in terms of states, for the moment. I am more careful about this in appendix B.

[14]Note that I offer no theory about what sentences express what information. The hope is that I can be completely neutral on this issue; you can plug in whatever theory you think works. Note also that I don't intend to tie the Calibration Picture to the implicit folk theory of information behind normal English; for all I have said, the substitution instances of 's' could be sentences of an extended English that contains complicated, theoretical terms for individuating information.

[15]This account of belief, which involves an important distinction between the sentence used in the belief ascription and the state of the world that makes the belief true, and it with its emphasis on token belief-states of particular agents, is in keeping with the approach to belief ascriptions advocated by Mark Crimmins and John Perry in Crimmins & Perry (1989) and further developed by Crimmins in Crimmins (2002).

(i) if p obtains, it is true that s,

(ii) p obtains in c;

2. A has a belief-state b_s whose target is p;

3. b_s is well-enough calibrated in c;

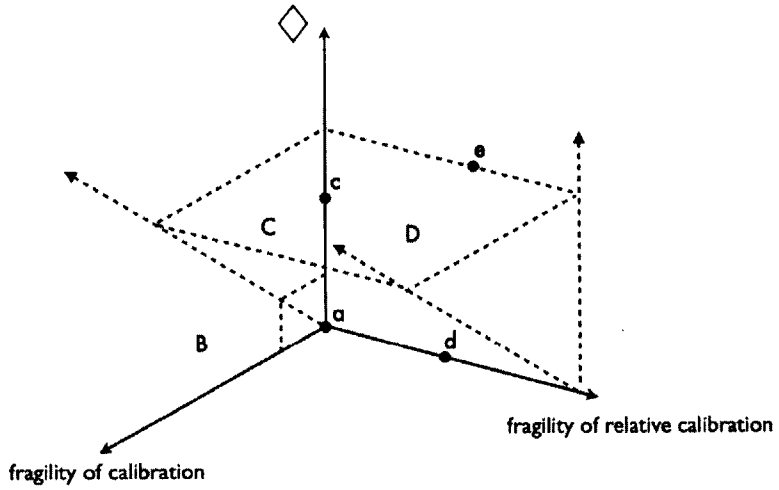
4. b_s's relative calibration relations are good enough in c.

> There is a spectrum of cases of learning. At one end are cases best described as ones in which the available information is available because of the internal state of the agent alone: those in which p obtains in all contexts, in which b_s is well calibrated in all contexts, and in which the relative calibration relations of b_s are good in all contexts. As we deviate from that best case in any of these three ways, it becomes less and less apt to describe the information as available because of the internal state of the agent, more and more apt to describe the information as available because of the state of the agent's environment. There is no particularly natural place in the spectrum to draw a line between internally and externally available information; things are a matter of degree.

[Aside: In describing the spectrum of cases, I said that one end of the spectrum are cases in which p obtains in *all* contexts, and so on. When it comes to agents, as opposed to measuring devices, we are usually better off thinking of a more restricted set of contexts — in particular, all possible *future* contexts. The end of the spectrum containing cases most aptly described as ones in which information is encoded in the agent are ones in which p is true in all possible future contexts, in which the agent is perfectly calibrated in all possible future contexts, and so on. I discuss this more in section 3.5, where I talk about differences between measuring devices and agents that are relevant to the model.]

To get a better grip on the nature of the spectrum of cases, as construed by our SECOND PASS, consider the following diagram:

The diagram represents the space of all possible cases of learning. The space is three-dimensional; along each dimension the cases vary in one of the three ways in which the environment plays a role. The vertical axis, labelled '$\Diamond$', represents



the degree of contingency of the target state p. The higher up that axis a case is, the fewer contexts in which p obtains. The axis coming out of the page is labelled 'fragility of calibration'. It represents how fragile the calibration of the agent's belief-state b is with p. The further along that axis a case is, the fewer the contexts in which b is well-enough calibrated with p for learning to occur. The third axis is labelled 'fragility of relative calibration'. The further along that axis a case is, the fewer the contexts in which the relative calibration relations between b and other belief-states is good enough for learning to occur.

The origin of the space is labelled 'a'. It represents cases that are best described as the agent using information she's already got — when p is true in all contexts, calibration is reliable across all contexts, and relative calibration is reliable across all contexts. Paradigm cases of learning by inference typically approximate a.

The axes are not completely independent of each other. In particular, the contingency of p and the fragility of the calibration of belief-state b with p are not completely independent of each other — if p is necessary, then any b with

p as target counts as perfectly well calibrated. So not all points in the diagram correspond to cases that are really possible. In particular, the region below the plane that divides the cube in half along the diagonal, labelled 'B' in the diagram, does not correspond to real cases, for the reason just given. The really possible cases fill the other half of the cube above the diagonal plane.[16] So there's a line of cases along the 'fragility of relative calibration' axis that broadens as you go up the contingency axis.

There is a region above the diagonal inside the plane that contains both the contingency axis and the fragility of calibration axis; it is marked 'C' on the diagram. That region contains cases in which the relative calibration of the relevant belief-state is perfectly reliable, but either p is contingent, or the straight-up calibration of the relevant belief-state is less than perfectly reliable. Paradigm cases of learning from experience are typically in, or at least close to, the C region.

The rest of the half cube — that part not in C — is labelled 'D'. It contains all cases of learning in which relative calibration is less than perfectly reliable. This is where the interesting cases are — all the ones mentioned in section 1, for instance.

Along the '$\Diamond$' axis are cases in which p is contingent, but the straight-up and relative calibration of the agent's belief-state are perfectly reliable. I have labelled one such case 'c'. Maybe learning the content expressed by sentences like 'I am here now' or 'I exist' are c-like cases. Whenever an agent believes 'I am here now' it's true, and agents are not typically liable to believe anything inconsistent with 'I am here now' in any context.[17]

Along the 'fragility of relative calibration' axis are cases in which p is necessary and straight-up calibration is perfectly reliable, but relative calibration is not perfectly reliable. I have labelled one such case 'd'. The cleanest examples of anomalies for the Classical Picture are d-like. These include TEMPERATURE and FACTORIZATION.

[16]It doesn't really have to be *half* a cube. There's no representational significance to the angle of the intersecting plane. It just needs to pass through the 'fragility of calibration' axis, and the angle it makes with the bottom of the cube needs to be between 0° and 180°.

[17]Actually, I don't believe that agents are not typically liable to believe anything inconsistent with 'I am here now' in any context. But it's closer to true in this case than most. I discuss issues related to this point in section 4.

Cases like the one labelled 'e', in the plane containing the '$\Diamond$' axis and the 'fragility of relative calibration' axis, are cases in which straight-up calibration is perfectly reliable, but p is contingent and relative calibration is imperfectly reliable. Cases of learning 'contingent *a priori*' propositions, as they are called in the literature, are usually like this. So, for instance, when one learns that Julius invented the zip (to use a well-known example),[18] one acquires a belief that is true whenever you have it, as the referent of 'Julius' was fixed by Gareth Evans with the description 'the inventor of the zip'. But it is contingent that Julius, whoever he is, invented the zip. And you are liable to believe things incompatible with the fact that he did. For example. you are liable to believe that Whitcomb Judson did not invent the zip. But Whitcomb Judson is, in fact, Julius.[19]

The most useful thing about putting the Calibration Picture in the above way is that it isolates the part of the epistemic story that the Classical Picture ignores: the relative-calibration part. The Classical Picture is the picture one gets by restricting attention to the C region. The a cases *are* quite different from the rest of the cases in the C region, so if those are the only cases you are looking at, you will take a-case/other-case distinction to be very epistemically important. Classical distinctions are just ways of dividing a-cases from other C-region cases. But the Classical distinctions that divide a-cases from other C-cases do poorly at distinguishing cases in the D region, where relative calibration is imperfectly reliable. D region cases are not helpfully lumped with either a or non-a, C-region cases.

This is why TEMPERATURE is a particularly instructive example: it focuses attention on the relative-calibration part of the story.

3.5 The Difference Between an Agent and a Thermometer

Agents are more complicated than measuring devices, so it takes a little work to extend our relatively simple model of the calibration of measuring devices to cover agents and their beliefs. Let me explain.

For modelling purposes, there are two important differences between measur-

[18] The example is due to Gareth Evans; see Evans (1979).

[19] Actually, whether Whitcomb Judson is Julius or not depends on exactly how you define 'inventor'. It might be that Elias Howe is Julius. See http://en.wikipedia.org/wiki/Zipper.

ing devices and agents. Firstly, measuring devices indicate the state of a system at a particular time — roughly, the time of the measurement — whereas agent's beliefs can concern the state of the world at any time.[20] Secondly, on the conventional way of individuating a measuring-device's indicating states, the device can be in only one such state at a time, whereas agents can have many beliefs at a time. (Relatedly, the states a typical device indicates are incompatible with each other, whereas agents' beliefs can be compatible.)

There are existing models of agents' belief-states and of learning that take these special features into account; we can extended our calibration model by using the same techniques as the existing models. A very simple model of an agent's belief-states and learning that underlies other, more sophisticated models (like Bayesian models) is used in the semantics of epistemic logics introduced by Jaakko Hintikka:[21] the possible-worlds model. The possible-worlds model allows for the fact that agents can have beliefs concerning any time by characterizing the content of beliefs with whole *worlds*, whereas our simple model of calibration assigns *states* of the world as the targets of device states. The possible-worlds model allows for the fact that agent's can have compatible beliefs by representing beliefs with *sets* of possible worlds, whereas our simple model of calibration identifies the targets of device-states with *particular* system-states, which are mutually exclusive.

We can do this too; we can represent the targets and reflections of individual belief-states as sets of possible worlds. To characterise accuracy and reliability we'll need to assume a metric on worlds, and we need some way to extend that metric to sets of worlds. There are many ways to extend the metric to sets of worlds, and the choice between them is a little arbitrary. The one I opt for in the appendix is closely related to what mathematicians call the 'Hausdorff distance' between two sets of points in a metric space. The nice thing about my choice is that a set of worlds A is 0 distance from a set of worlds B if and only if $A \subseteq B$. So all the entailment relationships between sets of worlds are encoded in their distance relationships, as I define distance.

[20]Similarly, agents can have counterfactual beliefs, whereas measuring devices only indicate the actual state of the system. This point raises *a lot* of complications (not so much technical as conceptual complications); I ignore the point here, with the hope of shedding some light on the issue later by building on what I do here.

[21]I believe Hintikka (1962) is the classic text.

4 Calibration and Rationality

4.1 Calibration and the Agent's Total Belief-State

Rationality is primarily a matter of an agent's *total belief-state.* That's why standard models of agents like the possible-worlds model and the Bayesian model just are models of total belief-states. The constraints on rationality articulated using such models come in two kinds: constraints that apply to all total belief-states — things like 'total doxastic states that include both a belief in p and a belief in $\neg p$ are bad, whatever the p' — and constraints that past total belief-states impose on future ones — things like 'future total doxastic states ought to be compatible with with past ones'.[22]

So if we want to understand how rationality fits into the calibration picture — at least rationality of the kind the orthodox models are talking about — we need to relate an agent's state of calibration to her total belief-state.

Now, you might think this is a fairly trivial thing to do. A full specification of an agent's state of calibration includes a specification of all her individual beliefs (at each world, no less). Isn't that, by itself, a specification of the agent's total belief-state (at every world, including the actual one)?

In the special case when all the agent's beliefs are compatible, then it is indeed very straightforward to go from a list of the agent's individual beliefs to a characterisation of her total belief-state. But things are not so straightforward when the agent has incompatible beliefs. And it is exactly in the interesting cases that the calibration model was designed to model that this happens — i.e., when the relative calibration relations among the agents beliefs are imperfectly reliable. Let me explain the problem, and then solve it for you.

Recall our model of the calibration of a measuring device. On the one hand, we have the states of the device; on the other, the states of the system. Each device state d has a state of the system it is supposed to indicate, called it's *target.* Also, each d has some set of system-states such that, as a matter of fact, it is

[22]Call rational constraints of the first kind 'synchronic' and rational constraints of the second kind 'diachronic'. It has recently become a matter of dispute whether there really are any diachronic constraints, or that, instead, the appearance of such constraints is an artefact of the real constraints, all of which are synchronous. For an argument that the only rational constraints are synchronous constraints, see Hedden (2013); for a defence of diachronic constraints, see Carr.

compatible with the system being in one of those states that the measuring device be in *d*. The set of those system-states is called *d*'s *reflection*. The device state is accurate, relative to some background conditions, when its reflection under those conditions is close to its target; it is reliable when it is accurate over a wide variety of background conditions.

We can also model the relative calibration between two device-states; it will be entirely derivative on the straight-up calibration of each. Roughly speaking, the relative calibration between two device-states will be accurate, relative to some background conditions, when the relationship between their reflections, given those background conditions, is like the relationship between their targets (in a sense of 'like' we can be precise about if we want). It will be reliable when it is accurate over a large range of background conditions.

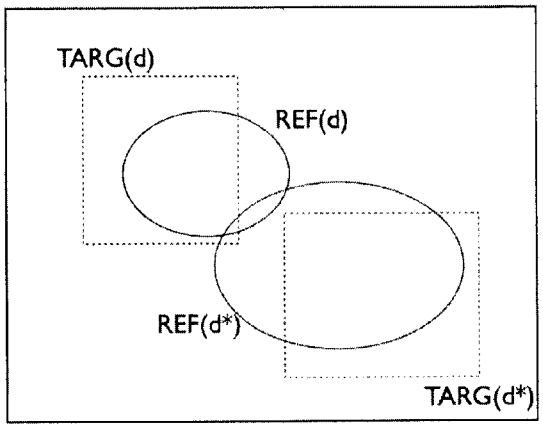
All of this goes for agents, just as it does for measuring devices. The relevant states of the agent are her individual belief-states. The relevant system is the whole world. Things are a little more complicated because the target states of the agent's belief-states are sets of world-states, rather than single states of the system, and they can overlap. It takes some technical jiggering to define things in the right way, given that complication. But the picture is basically the same.

[Aside: from here forward, I will just call individual belief-states 'beliefs". But don't let this mislead you; 'beliefs', in this sense, include belief-states that obtain only in counterfactual worlds. So an agent need not actually have a 'belief', in the sense I am using the term.]

We can represent the state of calibration of an agent pictorially. Here is a picture of the state of calibration of two beliefs.

The large black rectangle represents all of the different possible world states. The points in the rectangle represent maximal worlds-states — what traditionally are just called *worlds* in formal models like this. Regions in the rectangle represent less than maximal worlds-states, which you can think of just as sets of maximal world-states. The dotted squares inside represent the targets of two of the agent's beliefs. The whole-line ovals represent the reflections of those same beliefs. The red square and circle represent the target and reflection of one belief, the blue square and circle represent the target and reflection of another belief.

This agent is in an interesting situation, epistemically. The targets of her beliefs



do not overlap — they are mutually incompatible. But her reflections *do* overlap. So there are some states of the world in which the agent would have both beliefs, despite the incompatibility of their targets. That is, there are states of the world in which the agent would believe two incompatible things. That is, there are states of the world in which the agent would believe a contradiction.

This can never happen when an agent's relative calibration relations are perfect. The relative calibration relations between two of an agent's beliefs are perfect if and only if the relationship between the reflections of the beliefs is like the relationship between the targets of those belief-states, in a slightly technical sense of 'like'.[23] The important thing to know is that the overlap facts needs to be the same — the reflection of A is inside the reflection of B iff the target of A is inside the target of B; the reflection of A partially overlaps the reflection of B iff the target of A partially overlaps the target of B; and so on.

4.1.1 The Easy Case: Perfect Relative Calibration

When the relative calibration relations among *all* of an agent's beliefs are perfect, it is pretty clear how to go from a state of calibration to a characterisation of an

[23]For details, see appendix B.2.

agent's total belief-state, at any given world. For a given world w,[24] we can just look and see which of the agent's beliefs contain w in their reflection. The set of all the agent's beliefs that have w in their reflection will correspond to a particular state of the world — the (least specific) one that is in the *target* of all those beliefs. It will just be the intersection of all those targets. Call that state p. p is the agent's total belief-state, when the world is (in fact) in maximal-state w. If the agent is an epistemic superstar, then p and w will be identical; the less close p and w are, the worse off the agent is, epistemically speaking, when the world is in state w.

Norms of rationality are usually put in terms of the agent's total belief-state. A popular putative norm is that an agent's total belief-state ought not contain a contradiction; this amounts to saying there ought to be some p that the above procedure delivers, when applied to a rational agent. Equivalently, this amounts to saying that an agent ought to have perfect calibration relations. In section 4.3, I will argue that this is not so; but don't worry about that too much right now.

Another norm we can specify is this: an agent ought to be calibrated in such a way that, if at some time t_0, she takes the world to be p, and at some latter time t_1 she takes the world be q, then it ought to be the case that $q \subseteq p$. This and the previous assumption are both implicit in the standard possible-worlds model. By making this assumption, we can recover the standard possible-worlds model of our agent from our calibration model, in this special, perfect-relative calibration case. But we can make the one assumption without the other, and thereby define a broader class of agents than can be modelled in the standard way, if we like.

4.1.2 The Harder Case: Imperfect Relative Calibration

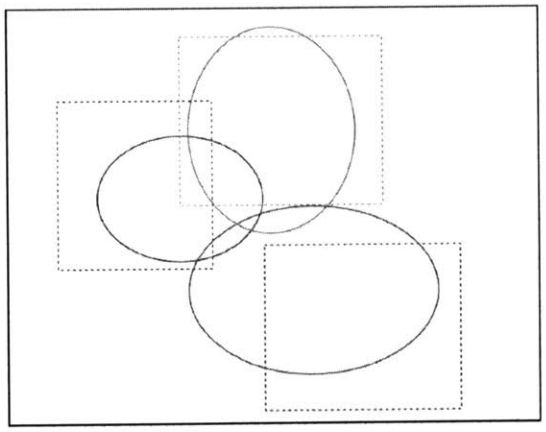
That's how things are when the agent has perfect relative calibration relations. But things are not so easy when our agent's relative calibration relations are imperfect. When an agent has imperfect relative calibration relations, there will be some states of the world w such that there is no p that is in the targets of all of the agents beliefs.

Maybe we can count the empty set as a state of the world — the null state — and say that *that* is the agent's total belief-state, in one of these problem

[24] And time t, strictly; agents have belief-states at a world at a time. I mostly ignore this in what follows, just for ease of exposition.

cases? We *could* do that, but it would be highly unhelpful. This involves treating all agents with imperfect calibration relations alike in worlds where they have incompatible beliefs, but there are obviously important differences among agents with incompatible beliefs, and it would be nice to be able to say something informative about these differences.

There is a natural way to generalize the procedure of section 4.1.1 so that it works for agents with imperfect calibration relations too. Here's the idea in pictures (the more official, technical statement of the idea is in the appendix). Step 1: Take the set of all the agent's beliefs. We can picture their state of calibration as follows.



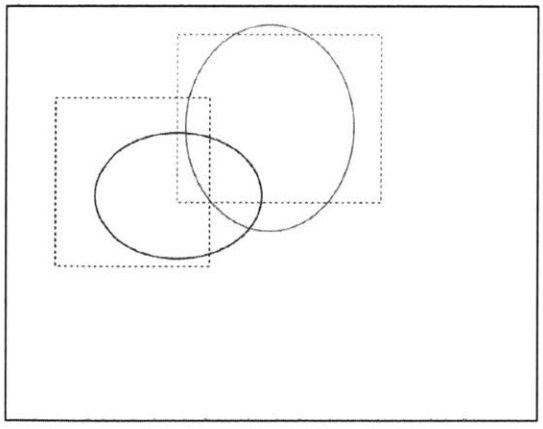
Just as before, the dotted squares represent targets and the whole-line ovals represent reflections, and like colours indicate which targets go with which reflections. Note that our agent is unrealistically simple – she has only three possible beliefs. This is to keep the picture manageable.

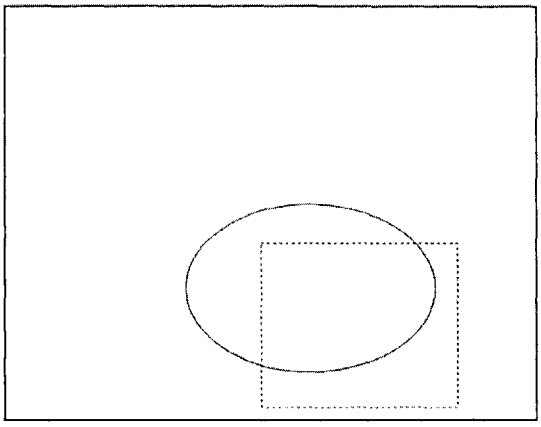
Step 2: take every maximal subset S of the agent's beliefs such that the relative calibration relations among those beliefs are perfect. Call such a subset a *fragment.* This term is an allusion to a way of modelling sub-optimal agents that has been part of the philosophy literature for a while.[25] The current approach is in the spirit

[25]See, for instance, Lewis (1982) and Stalnaker (1984).

of that work. But be careful; my use of the term is a little different. An agent can have multiple fragments, in my sense of the term, even if all her *actual* beliefs are perfectly compatible. It is enough that there be *some* world at which an agent has incompatible beliefs that she be fragmented, in my sense of the term. Obviously any remotely realistic agent will be fragmented in this sense.

A set of beliefs that looked like the one above will resolve into two fragments that look like this:



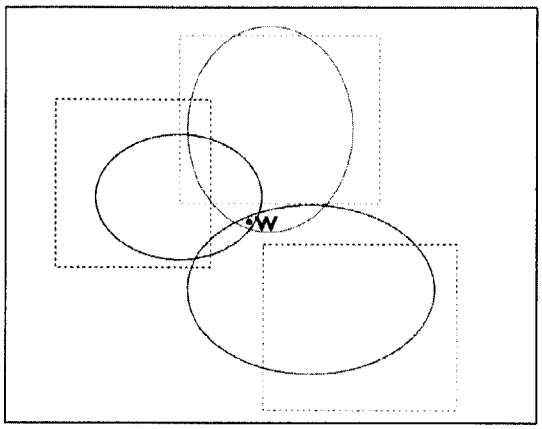


In general: the better the relative calibration relations among an agent's beliefs, the fewer fragments there will be. An agent with maximally good relative calibration relations will have just the one fragment, which will be the set of all her beliefs. On the other hand, an agent with maximally bad relative-calibration relations will have each of her beliefs in a fragment by itself.[26]

Step 3: give the fragments some order, which will henceforth be canonical. We'll give the fragments above the order in which they appear.

Step 4: take a maximal world-state w — say, the one labelled 'w' below.

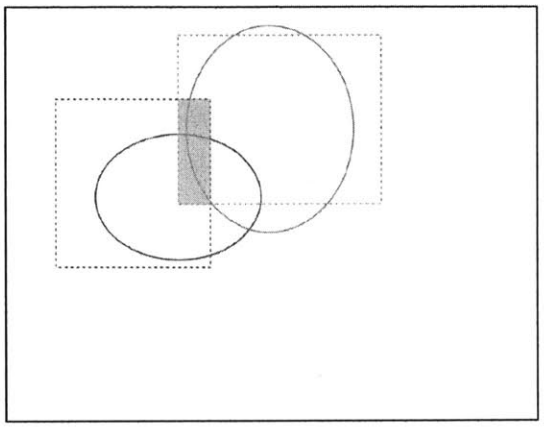
[26]Actually, it is doubtful there could really be such an agent, for reasons to do with the connection between the calibration of a belief and its content. But that connection is something I am silent on throughout this paper; I hope to have more to say about it in future work.

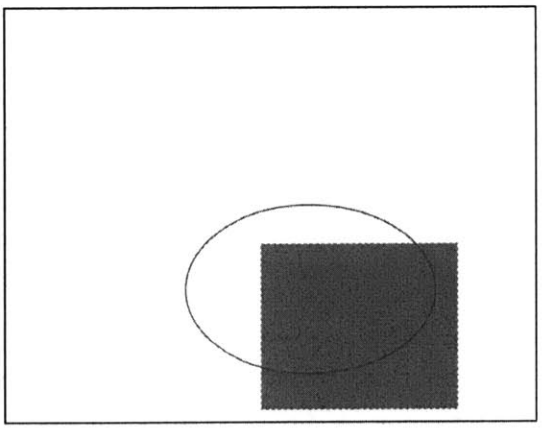


Take a particular fragment. See what belief-states in that fragment have w in their reflection. For our particular choice of w, both beliefs in fragment 1 have w in their reflections, and the lone belief in fragment 2 also has w in its reflection. This need not be so; there may well be many beliefs in a fragment that do not have w in their reflection, for a particular choice of w. It may be that *no* belief in a fragment has w in its reflection. That's fine. (We need some way to mark this; the official procedure in the appendix tells us that when there is no belief in a fragment with w in its reflection, Step 4 delivers the empty set.)

Now, define an individual fragment's total belief-state in the same way you define the belief-state of an agent with perfect calibration relations. That is: characterise the fragment's total belief-state with the intersection of the targets of all the belief-states in the fragment that have w in their reflection. **Note that this is the pivotal step**, where we move from considering reflections to considering targets.

Step 4 delivers the regions shaded below, for each of our example fragments:





Step 5: repeat step 4 for each fragment. Put the results in an ordered tuple — ordered in the same way as the canonical ordering of the fragments.

The end result of this will be an ordered set of states of the world. This ordered tuple represents the agent's total belief-state, when the world is in fact in state w. You can think of the agent as taking the world to be in each of the states in the tuple simultaneously. Of course, when the agent has incompatible

beliefs, the world cannot really be in each of these states simultaneously. But that shouldn't be a surprise; sometimes (most of the time) agents are wrong about what the world could really be like.

For each state of the world w there will be a tuple of world-states that represent the agent's total belief-state when w obtains. We can put norms of rationality in terms of these tuples, in a way that is analogous to the perfect relative calibration case. For instance, if we want to impose the diachronic constraint mentioned in section 4.1.1, we can do the following. Let us say that a tuple $\langle q_1, q_2, \ldots, q_n \rangle$ is a subset of another tuple $\langle p_1, p_2, \ldots, p_n \rangle$ iff $q_1 \subset p_1, q_2 \subset p_2, \ldots, q_n \subset p_n$. And, let us write $\langle p_1, p_2, \ldots p_n \rangle$ as $\mathbf{p}$. Then the analogue of the diachronic constraint on rational agents is this: iff at some time t_0, an agent's total belief-state is $\mathbf{p}$, and at some latter time t_1 her total belief-state is $\mathbf{q}$, then it ought to be the case that $\mathbf{q} \subseteq \mathbf{p}$. This more general constraint entails the special constraint for the case when the agent's relative calibration relations are perfect.

Just as a belief-state p can be represented as a set of possible worlds, a belief-state $\mathbf{p}$ can be represented as a set of ordered tuples of possible worlds, which we can write as $\mathbf{w}$. The set of all $\mathbf{w}$ is just the Cartesian product of n copies of sets of possible worlds, where n is the number of the agent's fragments. Call such a $\mathbf{w}$ a *notional world.* We can identify the maximally specific (really) possible states of the world w with tuples of the form $\langle w, w, \ldots, w \rangle$. We can call these the *metaphysically possible* notional worlds.

If we assume the diachronic constraint on rationality mentioned above, then total-belief-state models we have recovered are isomorphic to standard possible-worlds models; the only difference is that notional worlds play the role of possible worlds. We need not make the diachronic assumption, which gives us one way in which the class of total-belief-state-models defined by the above procedure is more general than the class of standard possible-worlds models. But it is more general in a second way: the standard possible-worlds model is only applicable to agents who have perfect relative calibration relations — i.e., to agents who would never have incompatible beliefs. Our first-personal models are applicable to agents with imperfect relative calibration relations.

Notional worlds are, in some ways, like what some theorists call 'logically possible worlds'. Putting things very vaguely, you can think of both notional

worlds and logically possible worlds as worlds that are possible in one sense, but not necessarily in another, metaphysical sense. But here is the difference.

Firstly, a set of notional worlds is indexed to an agent; it makes no sense to talk of notional worlds independently of an agent, as talking of notional worlds presupposes a particular agent in a particular state of calibration. Logically possible worlds aren't supposed to be like that; they are supposed to exist quite independently of agents, and the set of them does not change from agent to agent.

Secondly, it is clear in what sense a particular merely notional world is possible, and in what sense it is not. It is 'possible' only in the sense that there is some (really possible) way for a particular agent to be such that the agent's total belief-state is well characterized by that notional world. It is well-characterized that way because of the relationship of the beliefs she has, when she is that way, to real possibilities. A merely notional world is *not* literally possible at all.

On the other hand, it is completely unclear in what sense a merely logically possible world is possible, and what metaphysical possibility is, if it isn't something that applies to all real possibilities.

4.2 Simple Impossible Beliefs

Notional worlds give us theorists, who know the real deal when it comes to modal reality, a way to say something useful about the total belief-state of an agent who is misguided as to what is truly possible, and as a result has a total belief-state that has no real possibility as its object. It does this by construing the agent's total belief-state as a complex thing, involving the simultaneous believing of more than one real possibility.

Something that is *not* modelled using notional worlds is an agent who has a simple, un-complex belief in an impossibility — a belief that has no real possibility as its object and also cannot be construed as involving the combination of incompatible (but real) possibilities. At least, it has no way to distinguish among agents like that; all such agents' total belief-states get modelled with the null-state.

Is this an unfortunate limitation of the model? That's actually a complicated question.

You might think the answer is obviously 'yes', because you think that there are,

obviously, important distinctions to be made among agents who have beliefs like that — beliefs that have no real possibility as their object and are not complexes of real possibilities. To make your point, you might draw my attention to agents that we would use the following belief-ascriptions to describe:

- 'Anne believes that some bachelors are married'.
- 'Adam thinks gold has atomic number 80.'
- 'Elena thinks Fermat's Last Theorem is false.'

Surely, you will say, such agents' have beliefs that, firstly, have no real possibility as their content; secondly, are not complexes of real possibilities; and thirdly, are importantly different from each other.

If that's what you think, then here is where I agree with you: I agree that these agents are importantly different. And I agree that these agents are, in some way, going wrong with what is really possible. And I am happy to concede that there is probably some clear, rigorous way of using the words 'content' such that it is true to say that these agents have beliefs with 'different impossible contents'.

But here's what I don't think: there is some special way the world could be, an impossible way, that these agents represent the world as being. That commits *me*, an omniscient theorist, to an impossible possibility, and I have no idea what that *is*, let alone believe there are such things.

So however you get the result that these agents have beliefs with 'different impossible contents', I am going to want to know: how can I translate your theory into a theory about the relationship between the belief-states of the agent and the possible states of the world — i.e., really possible states of the world?

Now, I strongly suspect that any plausible translation scheme from your theory to a theory that invokes only real possibilities is going to go by way of notional worlds — i.e., by way of complexes of actual possibilities. My general reason for this suspicion is something like this: think about what, exactly, the world needs to be like, according to the agent, for one of the above belief-ascriptions to be true of her. Take Anne, for instance. What exactly does Anne think the world is like? Don't tell me 'she thinks that the world contains an unmarried bachelor'; that is

completely uninformative, to me who knows what is really possible and what is not. I wouldn't have asked if I thought that was a satisfactory answer.

There's got to be some sense in which she has a belief about bachelors — i.e., about things that are necessarily unmarried. And there's got to be some sense in which she believes, of one these things, that they are married. It is hard to see how to fill in the details in a way that makes Anne's picture of the world nice and unified.

Maybe what's going on is that Anne in one context is liable to affirm that Matthias is a bachelor — i.e., that he is unmarried — but in another context she will say of Matthias that he is married (maybe Matthias is in disguise, and part of the disguise is a wedding ring and many a mention of his partner). If that's what Anne is like, then obviously the way things seem to her is well captured by a notional world.

I think this point is completely general; if you try to fill out what an agent's picture of the world is like, when a belief-ascription like the one above is true, you will find yourself appealing to something like a notional world. And that makes me think that whatever theory of content you have that enables you to distinguish between Anne and Adam and Elena will, when you translate into real-possibility-talk, involve notional worlds.

You do not have to dig very deep to see the dis-unified nature of Anne's world-view. In some cases you will have to dig deeper. Take Adam. Adam is a less strange case than Anne. I can relate to Adam pretty well. Had Wikipedia told me that gold has atomic number 80, I would have believed it too.

But what, exactly, would I have been believing? Something like this: that if you zoom in on a piece of gold, you'll find that it's atoms have 80 electrons orbiting around them, and 80 protons and 80 neutrons in the centre. Though, of course, if that were true, then the thing you zoomed in on would not be gold. So I would have believed that gold is not gold. What?

Maybe this is what is going on: among other things, I am willing to affirm that a particular piece of gold is, in fact, gold. I am also willing to affirm that a particular piece of gold has atomic number 80 — i.e., is not, in fact, gold. Clearly, this is a situation in which my belief-state is well captured by a notional world.

Elena's case is the hardest to fit into the notional world framework. It is

the hardest to model using genuine possibilities in any way at all. That's why philosophy of mathematics is hard; it is not at all obvious how we should construe the content of mathematical beliefs. I do not have a lot to add to the literature on this topic. I do insist that if we, as theorists of inquiry, are to make sense of Elena and agents like her, we need to find some way of relating her mental state to world-states — i.e., to genuinely possible world-states.

My suspicion, for what it's worth, is that a belief-state like Elena's is correctly modelled in terms of relationships between other belief-states that are more directly connect to world-states. What it is for Elena to believe that Fermat's Last Theorem is false is, on this model, for her total belief-state to meet certain higher-order, structural conditions. This means that there will be a lot of variability as to what notional worlds are used to model any particular agent with a belief like Elena's. No particular notional world get rules in or out just in virtue of believing that Fermat's Last Theorem is false, if this suspicion is right. It is only in combination with the rest of the agent's beliefs that we can connect the agent's mathematical beliefs to states of the world.

4.3 Ideally Rational Agents

We have seen one constraint on rational calibration that is often made: that, if the agent's total belief-state at t_1 is $\mathbf{p}$, and at t_2 later than t_1 it is $\mathbf{q}$, then $\mathbf{q} \subseteq \mathbf{p}$. What other constraints might there be?

Until about the '70s, many people thought that an ideally rational agent had to be a non-fragmented agent — i.e., ideally rational agents never made mistakes about what propositions were compatible with what, and so their total belief-states were always characterizable by a set of notional worlds that are really possible. But Saul Kripke taught us that this is *not* a plausible constraint on rationality.[27] It is not a failure of rationality, in any interesting sense, to be ignorant of that fact that Hesperus is Phosphorous, or that water is H_2O, or that the atomic number of gold is 79. But to be ignorant of these facts is to mistakenly think that Hesperus rising is compatible with Phosphorous setting, or that there being water in the cup is compatible with there being no H_2O in the cup, or that striking gold is

[27]He did this in Kripke (1980).

compatible with striking the element with atomic number 78. And an agent with these mistaken beliefs is an agent with imperfect relative calibration relations – i.e. is a fragmented agent.

This lesson has, in general, been accepted. But many hold out hope for a close surrogate view. The idea is not that the ideally rational agent's set of notional worlds must be really possible. It is instead that there is some other special set of notional worlds that are distinctive of the ideally rational agent – a set which contains the really possible worlds as a subset. On this view, there is some state of relative calibration that is distinctive of ideal rationality, so all possible ideally rational agents are calibrated in this way and, as a result, are associated with the same set of notional worlds.

This view has got its most explicit and fullest expression in the work of David Chalmers;[28] he calls the special notional worlds 'epistemic possibilities'. But the view is implicit whenever people assume that there is some well defined set of propositions that are knowable *a priori* to all ideally rational agents. If there is such a well defined set, it is exactly the propositions that are true and not false at all the special notional worlds. Philosophers assume there is such a set of propositions *all the time.*

I say there is no such special set of notional worlds. Here is my argument. Suppose that there were such a special set. Then there would be some notional world **w** outside the special set such that, to have a perspective characterized by a set of notional worlds that includes **w** is *ipso facto* to be irrational. But, I say, there is no such **w**. I say that for any given notional world, constructed from any given state of calibration, you can come up with a situation in which failing to rule out that notional world (and, hence, having a total belief-state characterized by a set that contains that notional world) is clearly *not* a failure of rationality.

Here, in general, is the reason why (this will seem a little vague; it will become clearer what I mean as we go). For any given state of the world, there are always many ways of having a doxastic attitude concerning that state of the world. For any two states of the world, you can always find a way of having a doxastic attitude

[28]The main task of Chalmers (2012) is to develop a version of this view. Note that Chalmers *also* has the further ambition of holding on to some form of the pre-Kripke view, though this is not argued for in Chalmers (2012).

concerning the one that is epistemically independent of some way of having a doxastic attitude concerning the other. By exploiting this independence, one can always come up with a situation in which an agent believes that a certain state obtains, and yet fails to rule out another, incompatible state, and this involves no failure of rationality — the independence of the ways in which these doxastic attitudes are had means that there is no clue of the agent's imperfect relative calibration that is accessible to the agent herself. That being so, for any given merely notional world, one can always come up with a situation in which an agent fails to rule out that notional world, and this is not a failure of rationality.

To see what I have in mind, let's work through some examples. First, a warm up case. Take a state of the world in which there is water in a particular cup at particular time and an incompatible state of the world in which there is no water in that cup at that time. Saul is a perfectly rational fellow with as much evidence as the best scientists of the seventeenth century had concerning the chemical composition of water. He looks at a cup of water. There is a clear sense in which he believes there is water in the cup. For instance, if you were to ask him if there was water in the cup, he would say 'yes'. There is also a clear, independent sense in which he fails to believe that there is water in the cup. If you were to ask him if there were H_2O in the cup, for instance, he would say 'I have no idea'. So the merely notional world in which there is water in that cup at that time and not water in that cup at that time is *not* a notional world that one must rule out to be rational.

Warm Up Case 2: take a state of the world in which London is pretty and an incompatible one in which London is not pretty. Pierre is a perfectly rational fellow with excellent evidence provided by his French comrades that London is pretty. He believes them; if you were to ask him "Est-ce que Londres est jolie?", he would say 'oui'. So he believes London is pretty. Pierre also has excellent evidence, gathered from his time living in London, that London is not pretty. Were you to ask him "Is London pretty?" he would say 'no', and shudder a little bit. So he also believes London is not pretty. *A fortiori,* he fails to rule out that London is not pretty. So the merely notional world in which London is pretty and London is not pretty is *not* a notional world that one must rule out to be rational.

Both Saul and Pierre's fragmented beliefs are naturally described using two

different terms that have the same semantic value rigidly; for Saul, the terms are 'water' and 'H_2O'; for Pierre, they are 'London' and 'Londres'. This provides a clue to the general formula for finding a case where it is rational to have one's belief-state characterized by a given merely notional world, involving two incompatible states of the world. What you do is find a situation in which the agent is well described as believing the one state obtains using one rigid term, and well described as failing to rule out that the other state obtains using another rigid term with the same semantic value. You can always find such a situation because being apt to be described as believing something (or failing to not believe it) using a rigid term is *easy*; there are a million ways it can happen. You then fill in the details in a way that makes the agent's belief-state rational. You can always fill in the details in this way because believing something in the rigid-term-1 way is, in general, independent of believing something in the rigid-term-2 way, in the sense that believing something in the rigid-term-1 way is no guarantee of an epistemological upshot for the agent — something she can use to figure out that her belief is incompatible with her rigid-term-2 doxastic state. This is the interesting feature of rigid terms: believing something in a rigid-term way does not constrain your psychology, vis-a-vis the semantic value of the rigid term, in any substantial way.

There are two theses that need to be true, for the above formula to be fully general:

ASCRIPTION: For every pair of incompatible states of the world p, q, there is a possible agent A such that

1. there is a doxastic-attitude ascription s_1 that is true of A, that ascribes a belief with p as target, and that contains a term that has semantic value v rigidly, and
2. there is doxastic-attitude ascription s_2 that is true of A, that ascribes a failure to rule out q, and that contains a term that has semantic value v rigidly.

RATIONALITY: If, for given p, q, there is a possible agent A and doxastic attitude ascriptions s_1, s_2 that satisfy the above, then for p, q there is a possible *rational* agent

who satisfies the above.

I think RATIONALITY should be uncontroversial. You've given the game away, once you've admitted that there is *an* agent who believes that there is water in the cup but no H_2O in the cup, assuming those do indeed have the same semantic value, or that there is an agent who believes that Londres est jolie but that London is not pretty; the details are too easy to fill in as one pleases. The more interesting question is whether ASCRIPTION is true.

What would a counter-example to ASCRIPTION look like? It would have to involve a v such that there was at most one way to believe things about v rigidly; there would have to be at most one rigid-ish doxastic route from the agent to v. Is there anything like that, in the world?

Such a v would have to be a very strange thing. For most objects in the world, there is no barrier to thinking about them rigidly by the mental equivalent of rigidifying a definite description, and there are obviously lots of definite descriptions that pick out any one thing — 'the thing that was on the table on Tuesday', 'Dan's favourite thing', and so on. So for any normal object, there is no barrier to thinking about it rigidly in lots of ways. So if there is a v that forms the basis of a counter-example to ASCRIPTION, it's got to be something that is very hard to think about; you need some very special doxastic relationship with the thing.

The best candidates for such a v are things that are tied very closely to the agent's own perspective — things like her own sense data, of experiences, or qualia, if there are such things. Russell's theory of acquaintance[29] is the paradigm case of a theory on which things like sense-date are exactly the v we are looking for: they can only be believed about rigidly in one special way. This is no coincidence; it was precisely the kind of epistemic ambitions of my opponent that (among other things) motivated Russell's theory. But Russell was wrong; there is no v that one can think about rigidly only one way. If sense data exist, they are not like that.

This is obvious if you are a physicalist; if any particular experience just is a particular brain being in a particular state, then obviously one can believe one's self to be having an experience and yet fail to rule out that one is not having that

[29]See Russell (1985).

experience — i.e., fail to rule out that one's brain is not in the relevant state. But even if you aren't a physicalist, you shouldn't believe experiences, or sense-data, or whatever, are involved in an exception to ASCRIPTION. An agent can, for instance, believe they are having sensation S, where 'S' is a rigid term the agent introduces to refer to her current sensation, but fail to rule out that they are not having sensation S', where 'S'' is a rigid term the agent introduced yesterday for, as it happens, the very same sensation.[30]

Moral: there are no counter-instances to ASCRIPTION. So it is true. And, since it and RATIONALITY are true, there is no notional world such that failing to rule out that notional world makes one *ipso facto* irrational. So there is no state of calibration that is distinctive of all rational agents.

4.4 Minkowski Epistemology

If there were a special state of relative calibration that was distinctive of the ideally rational agent, we could use that state of relative calibration to define a clear distinction between the kind of learning that occurs when an agent acquires new information from her environment, and the kind of learning that occurs when an agent uses information she has already got. Acquiring new information would occur whenever an agent acquired warrant to rule out one of the special notional worlds that are distinctive of the ideally rational agent. Indeed, it would be convenient to represent information just as the set of special notional worlds consistent with that piece of information; ideally rational belief-update would involve moving from a belief-state to the intersection of that belief-state with the acquired information, represented as a set of notional worlds.

Using information an agent already had would occur whenever an agent acquired warrant to rule out a notional world outside the special set. Indeed, given an agent's total belief-state, the ideally rational belief-state for her to have would just be the intersection of her current belief-state and the set of special notional worlds. This would give us a definite answer to the question of what it is rational

[30]The use of 'S' here is an allusion to Wittgenstein's Private Language Argument - in particular, the example of the diarist. Personally, I think one of the things Wittgenstein was getting at in the Private Language passages of Wittgenstein (1953) is that even sense-data are not invulnerable to the kind of error distinctive of imperfect calibration.

for an agent to believe at time, given her total belief-state at that time.

Another clear distinction that comes along with the distinction between acquiring information and using information one has already got: the distinction between things that can be known *a priori* and things that can only be known *a posteriori*. Something is knowable *a priori* if it is knowable using information an ideally rational agent has prior to acquiring any information from her environment. The *a priori* knowable truths would just be those truths that are compatible with every special notional world.

A final distinction that comes with the special set of notional worlds: the distinction between irrationality and ignorance. An agent would be ignorant to the extent she had failed to rule out special notional worlds other than the actual world. She would be irrational to the extent she had failed to rule out non-special notional worlds.

But there is no special state of calibration distinctive of the ideally rational agent, and, hence, no clear distinction between irrationality and ignorance. This distinction is pretty fundamental to epistemology, as it is traditionally conceived. If we can't make that distinction, how are we supposed to do epistemology?

Here is my proposal.

We hold on to the distinction between ignorance and irrationality. But we introduce a new relativity to the notion; the difference can only be drawn relative to a choice of relative calibration relations, and, hence, a choice of notional worlds.

The relationship between my way of treating the ignorance/irrationality distinction and the traditional way is analogous to the relationship between the way special relativity treats the space/time distinction and the way that distinction is treated in classical mechanics. Special relativity tells you what the fundamental, invariant thing is: Minkowski space-time, which consists of events and their space-time relations. And it tells you how to decompose these space-time relations into spatial relations and temporal relations, relative to a frame of reference. The thing that plays the role of Minkowski space time in my proposal is the state of calibration of the agent, and her distance from perfect calibration; it's the fundamental, invariant thing. The thing that plays the role of a frame of reference in my proposal is a state of relative calibration, or equivalently, a set of notional worlds. Note that this frame-of-reference state-of-relative-calibration need not be,

and for most interesting purposes will not be, the state of relative calibration of the agent herself. With a particular state of relative calibration in hand, we can then decompose the agent' distance from perfect calibration into a part of that distance due to ignorance and a part due to irrationality.

Just to be clear: there are three things in play. There is, firstly, the set of all notional worlds of a given arity n (the arity is determined by the degree of fragmentation of the relevant agent). It is just the n-ary Cartesian-product of the set of all possible worlds. Secondly, there is the subset of those notional worlds that characterizes the agent's total belief-state. Thirdly, there is the frame-of-reference set of notional worlds, against which one can decompose the agent's total belief state into a state of ignorance and a state of irrationality.[31]

There is a spectrum of sets of frame-of-reference notional worlds to choose from. At one end is just the set of really possible worlds. Any fragmentation counts as irrationality, relative to this set of notional worlds. For many purposes, this choice of notional-world-set will do just fine. You need a relatively sophisticated agent, and relatively subtle situation, before this becomes an unhelpful choice; hence the hundreds of years it took anyone to notice that what is necessarily true does not coincide with what can be known *a priori*.

Further along the spectrum are larger and larger sets of notional worlds. More and more counts as ignorance, and less and less as irrationality, as you move along the spectrum.

It is not obvious that there is any other end to the spectrum. That would require some upper limit to how fragmented an agent can get. And that in turn depends on whether there is some upper limit to how many different ways there are to have beliefs concerning the one thing. I highly doubt there is some upper limit; I suspect that, for any possible fragmented agent, there is a possible more fragmented agent.

For any given agent, in any given situation, what determines which set of notional worlds we should pick to make the irrationality/ignorance distinction

[31]Note that there is some slight technical trickiness here because of a possible mismatch between the arity of the agent's notional worlds and the notional worlds one wants to use as a frame-of-reference; the solution is to fill out the smaller arity notional worlds with more copies of worlds that are already in them.

against? We can have different theories about this. The options here parallel the options for broadly contextualist theories of knowledge: straight-up contextualism, relativism, sensitive invariantism, and expressivism. Any choice one makes here, about rationality, fits well with the same choice for knowledge. Personally, I favour the expressivist option (which, following Hartry Field, I take to be the same as the relativist option, properly understood).[32] I think the choice of notional worlds depends on *us* as theorists, and our purposes in modelling the relevant agent. It is a practical question to be answered by the theorist, rather than a factual question to be answered by what the agent modelled is like. We already know what the modelled agent is like: she has a certain state of calibration.

Here are some examples of how a choice of notional worlds might be guided by the practical concerns of the theorist.

One practical aim you might have in making the distinction is remedy: you might want to figure out what the agent should do to bring herself closer to believing all and only the truths. Should she be doing a lot of sitting and thinking, or should she be doing a lot of experiments, or what? If one is lucky, one can choose a set of notional worlds so that the distinction between irrationality and ignorance relative to that set lines up roughly with the distinction between more sit-and-think kinds of remedies and more do-experiments kinds of remedies. But note two things. Firstly, the remedies won't always divide so neatly; as was noted in Part I, there are going to be many cases where the right remedy to an imperfect belief-state is not going be very close to either sitting and thinking or doing experiments, but will look more like something in the middle. Secondly, if this is your purpose, then your choice of notional world is going to be highly contingent on the particular nature of the agent in question. What agents are in a position to work out by sitting and thinking — in a real, practical sense of 'in a position to work out' — is very different from one agent to another.

Another practical aim one might have as a theorist is being able to understand communication between several agents. You will want to choose your notional worlds in such a way that the utterances of the agents carry new information. For instance, you may be trying to understand a conversation between Holmes and

[32]Field (2009).

Watson concerning whether or not the clues entail that the butler did it. One will get a more perspicuous account of what is going on if one models things in such a way that Watson is ignorant, rather than irrational, in not knowing that the clues entail that the butler did it.

I will have more to say about the link between calibration and communication in forthcoming work.

4.5 An Application: Frontloading Arguments

Here is an argument, variations of which have been used in many places in the literature[33] to argue that agents are in a position to have quite a lot of knowledge *a priori*. Let H be some proposition an arbitrary agent A is in a position to know.

P1 A is either in a position to know H *a priori*
or on the basis of empirical evidence E.

P2 If A is in a position to know H on the basis of empirical evidence E, then she is in a position to know that if E, then H *a priori*.

C A is in a position to know either H or that if E, then H *a priori*.

As H was arbitrary, this argument works for all H. So it appears we are in a position to know quite a lot *a priori*.

What is a calibrationist to make of this argument? Calibrationist that I am, I can discern three different ways of understanding the phrase '*a priori*' in the frontloading argument. On the first of these three, I reject premise 1. On the second, I reject premise 2. On the third, I accept that the argument is sound, but the conclusion is not nearly as exciting as the *a priori* enthusiast might have hoped. Let me explain.

Reading 1: the phrase '*a priori*' presupposes that there is a special set of notional worlds that characterize the state of calibration of an ideally rational agent, for reasons articulated in section 4.4. There is no such set. Now, this does not by itself mean we can just ignore the premises; often a sentence can manage to assert some relatively interesting content despite presupposition failure. But

[33]See, for example, chapter 4 of Chalmers (2012), White (2006), Wright (2002), Schiffer (2004) and the 'Explainer' example in Hawthorne (2002)

anything in the neighborhood of P1, on this reading of '*a priori*', that might survive presupposition failure is certainly false. As I said way back at the beginning of Part 1, there are many things many agents are in a position to know that they aren't in a position to know in a particularly *a priori*-like way, nor in a particularly empirical-evidence-like way.

Reading 2: 'is in a position to know *a priori*' means 'is in a position to know in any environment'. 'Is in a position to know on the basis of empirical evidence E', then, means something like 'is in a position know in an environment in which E is true (and that does not include all environments)' On this reading, P1 is certainly true. One should note that the instances of H which make the first disjunct of P1 true will be vanishingly rare for any remotely realistic agent, but that's ok. On the other hand, P2 is certainly false. All the examples that I used to motivate the calibration picture back in Part 1 provide ready counterexamples. Damien is in a position to know that 55°F is about 13° in environment that are, in fact 55°F (or in whatever environment elicits the 55°F judgment, in the alternative brain-surgery case). He is not in a position to know in any environment that if it is 55°F, then 55°F is about 13°. And it is hard to see what even *prima facie* reason there is to suppose otherwise.

Reading 3: 'is in a position to know *a priori* means 'is in a position to know on the basis of information every agent has available *relative to such-and-such state of relative calibration* — where such information is whatever is true and not false at every notional world characteristic of the relevant state of calibration. If that's what you mean, then I am all for your front loading argument. I accept your premises and your conclusion.

The conclusion of the reading-3 version of the argument is a *somewhat* interesting result — it shows an interesting fact about how information-relative-to-a-state-of-calibration works in general. But it is not as exciting as you might have hoped. Something that doesn't follow from it, for instance, is that if only an agent were smart enough, she could work out lots of interesting things without having any experiences. In fact, almost nothing follows from the conclusion of the reading-3 version of the argument concerning what any actual agent might be able to know in an actual situation. This is because the sense in which the relevant information is 'available', on the reading 3 version, can float completely free of the

particular agent's skills and capacities, and the ways in which her environment is conducive to knowledge acquisition, for that agent.

5 Conclusion

The Classical Picture is bad, but it was not obvious what alternative picture of learning there is. Above, I have tried to provide one. With a clear alternative in place, we are in a better position to see how to carry out epistemology without relying on Classical, absolute distinctions between information the agent already has and new information, between experiences that justify and experiences that enable, between *a priori* and *a posteriori* knowledge, and between irrationality and ignorance. I think this is progress.

A Appendix I: Calibration of Measuring Devices

A.1 Accuracy and Reliability

Let the set of device-states be D. Let the set of world states be W. Let the set of system states S be a partition on W. Let there be a metric σ on S. Let $\mathsf{TARG}(d) \in S$ be the state that $d \in D$ is supposed to indicate well — i.e., d's *target*. Let $\mathsf{REF}_b(d)$ be the state the world is in when the device is in d and background conditions b obtain. You can think of b as a state of the whole of the rest of the world, beyond device and system. I will refer to $\mathsf{REF}_b(d)$ as d's *reflection* (relative to b).

There are two senses in which d can indicate $\mathsf{TARG}(d)$ well: it can be *accurate* and it can be *reliable*. Device state d is accurate relative to b iff $\sigma(\mathsf{TARG}(d)), \mathsf{REF}_b(d))$ is small. In words: d is accurate relative to b iff the state the system is in, when the device is in d and background conditions b obtain, is close to the state d is supposed to indicate.

State d is *reliable* iff the range of background conditions b for which the state indicates accurately is large.

Let $\mathsf{REF}(d) = \bigcup_b \mathsf{REF}_b(d)$; call it d's *reflection* (*tout court*). A good measure of how good the overall calibration of d is how close $\mathsf{REF}(d)$ is to being inside $\mathsf{TARG}(d)$, where this is measured in terms of σ.

The device as a whole is accurate iff its state are, for the most part, accurate; it is reliable as a whole iff its states are, for the most part, reliable.

A.2 Relative Calibration

Let d, $d^\star$ be two device states whose targets are in the same system — i.e., are such that $\sigma(\mathsf{TARG}(d), \mathsf{TARG}(d^\star))$ is defined. How good the *relative calibration relative to background conditions* b is, between d, $d^\star$, is inversely propositional to the size of $|\sigma(\mathsf{REF}_b(d), \mathsf{REF}_b(d^\star)) - \sigma(\mathsf{TARG}(d), \mathsf{TARG}(d^\star))|$. When $\sigma(\mathsf{REF}_b(d), \mathsf{REF}_b(d^\star)) = \sigma(\mathsf{TARG}(d), \mathsf{TARG}(d^\star))$, the relative calibration of the states relative to b is perfect. The more $\sigma(\mathsf{REF}_b(d), \mathsf{REF}_b(d^\star))$ differs from $\sigma(\mathsf{REF}_b(d), \mathsf{REF}_b(d^\star))$, the worse the relative calibration between d, $d^\star$, relative to b

The relative calibration between d and $d^\star$ more generally is a function of $\sigma(\mathsf{TARG}(d), \mathsf{TARG}(d^\star) - \sigma(\mathsf{REF}_b(d), \mathsf{REF}_b(d^\star))$ for all b. The bigger the sum of these differences, the worse the relative calibration between d and $d^\star$.

B Appendix II: Calibration of Agents

B.1 Accuracy, Reliability for Agents

Let W be the set of maximal ways for the world to be throughout time — i.e., the things that are usually called 'worlds' in formal models. Let ω be a metric on W.

Let $\overline{W}$ be the power set of worlds; we identify each member $p \in \overline{W}$ with the proposition that is true at each of its members. We need to extend ω to $\overline{W}$. There are numerous ways of doing this. A way that is particularly useful for our purposes is the following. First, we define the distance between a member of W and an arbitrary $q \in \overline{W}$ like this:

$$\omega(w, q) = \inf\{\omega(w, w') | w' \in q\}$$

—i.e., the distance between w and q is the shortest distance between w and a member of q. Now we define the distance between two arbitrary members $p, q \in \overline{W}$ as

$$\sup\{\omega(w, q) | w \in p\}$$

— i.e., the greatest distance between a member of p and q, as defined above.

To summarize:[34]

$$\omega(p, q) = \sup_{w \in p} \inf_{w' \in q} \omega(w, w')$$

Note that ω, defined this way, is *not* a metric, in the technical sense — there are distinct elements with distance 0, and ω is not symmetric.[35]

[34]This definition of distance between sets of points is close to what is called 'the Hausdorff distance'; the only difference is that the Hausdorff distance between p and q is the maximum of $\omega(p, q)$ and $\omega(q, p)$. This last step is needed to turn distance as I've defined it into a proper metric.

[35]However, $\omega(p, p) = 0$ for all p, and ω obeys the triangle inequality, so it is what is called a 'pseudoquasimetric'.

The nice thing about this extension of ω to $\overline{W}$ is that, given the definition, $p \subseteq q$ iff $\omega(p, q) = 0$. Hence p entails q iff $\omega(p, q) = 0$. You can think of $\omega(p, q)$ as a measure of how close p is to entailing q — the lower $\omega(p, q)$, the close p is to entailing q.

Let B be the set of belief-states of the agent. Let $\mathsf{TARG}(b) \in \overline{W}$ be the target state of b — i.e., the p that goes with b in the SECOND PASS of section 3.4. More simply: $\mathsf{TARG}(b)$ is the set of worlds in which b is true. Let $\mathsf{REF}_c(b)$ be the set of worlds that are consistent with the agent being in state b, at some time or other, and with background conditions c. $W_{\mathsf{TARG}_c(b)}$ is b's reflection, relative to c.

Accuracy and reliability are defined in the way you would expect: b is accurate, relative to background conditions c, iff $\omega(\mathsf{REF}_c(b), \mathsf{TARG}(b))$ is small. In words: b is accurate relative to c iff the propositions that is true when the agent believes b and background conditions c obtain is close to entailing that b is true. Belief-state b is reliable iff the range of background conditions over which it is accurate is large.

Let $\mathsf{REF}(b) = \bigcup_{\forall c} \mathsf{REF}_c(b)$; call it b's *reflection* (*tout court*). A good measure of how good the overall calibration of b is is how small $\omega(\mathsf{REF}(b), \mathsf{TARG}(b))$ is.

The agent as a whole is accurate iff her states are, for the most part, accurate; she is reliable as a whole iff her states are, for the most part, reliable.

B.2 Relative Calibration for Agents

Given the above definitions, the definition of relative calibration for agents beliefs states is as you would expect. It is stated slightly differently to relative calibration for measuring devices only because of the asymmetry of ω.

How good the $b^\star$-wise relative calibration relative, for a belief-state b, relative to background conditions c, is inversely propositional to the size of $|\omega(\mathsf{REF}_c(b), \mathsf{REF}_c(b^\star)) - \omega(\mathsf{TARG}(b), \mathsf{TARG}(b^\star))|$. When $\omega(\mathsf{REF}_c(b), \mathsf{REF}_c(b^\star)) = \omega(\mathsf{TARG}(b), \mathsf{TARG}(b^\star))$, the $b^\star$-wise relative calibration of b, relative to c, is perfect. The more $\omega(\mathsf{REF}_c(b), \mathsf{REF}_c(b^\star))$ differs from $\omega(\mathsf{TARG}(b), \mathsf{TARG}(b^\star))$, the worse the $b^\star$-wise relative calibration of b, relative to $b^\star$.

The $b^\star$-wise relative calibration of b more generally is a function of $\omega(\mathsf{REF}_c(b), \mathsf{REF}_c(b^\star)) - \omega(\mathsf{TARG}(b), \mathsf{TARG}(b^\star))$ for all c. The bigger the sum of these differences, the

worse the b^*-wise relative calibration of b.

The b^*-wise relative calibration of b will be perfectly reliable only if the agent treats b and b^* as having the inferential relationship they in fact have in all contexts (this follows from the way we defined ω). In general, the agents relative calibration relations will be perfectly reliable (if and) only if the agent treats all propositions as having the inferential relationships they in fact do have, in all contexts. In this special case (and only in this special case), an agent will have a consistent total belief-state in all contexts. Also just in this special case, whenever an agent comes to believe b, she will thereby come to believe all things entailed by the truth of b, and disbelieve all things incompatible with the truth of b. Hence, this special case is the case modelled by the orthodox possible-worlds model of an agent's belief-state and belief update.

C Appendix III: From State of Calibration to Total Belief-State

Start with an agent in some state of calibration at some time. Let B be the set of all the agent's possible belief-states. Let W be the set of maximal world states; let P be the power set of W. For each of the agent's belief-states $b \in B$, there is b's target, $\mathsf{TARG}(b) \in P$, and b's reflection, $\mathsf{REF}(b)$, P. These determine the relative calibration relations among the agent's belief-states.

Let a *fragment* F be a subset of B such that

(a) relative calibration relations among all members of F are perfect, and

(b) F is not a subset of any set that meets condition (a).

Let F_w be the set of all $b \in F$ such that $w \in \mathsf{REF}(b)$.

Let $\mathsf{TARG}(F_w) = \{x \in P : x = \mathsf{TARG(b)}\}$, for some $b \in F_w$. Let

$$p_{F_w} = \bigcap \mathsf{TARG}(F_w)$$

p_{F_w} is a member of P.

Let $\mathbf{F}$ be an ordered set of all $F \subset B$. Let $\mathbf{p_{Fw}}$ be the ordered set of p_{F_w} for all F and a given w, where the ordering corresponds to the order of members of $\mathbf{F}$.

For a given w, $\mathbf{p_{F_w}}$ is one way of characterizing the agent's total belief-state at w, at the relevant time, given her current state of calibration.

Here is another, better way. Let n be the cardinality of $\mathbf{F}$. Let W^n be the n-ary Cartesian product of W. Let $\mathbf{P}$ be the power set of W^n.

Let $\mathbf{q_w}$ be the largest member of $\mathbf{P}$ such that the union of all the first elements of $q_w = p_{F_{1,w}}$, the union of all the second elements of $q_w = p_{F_{2,w}}$, and so on.

For a given w, $\mathbf{q_w}$ is another way of characterizing the agent's total belief-state at w, at the relevant time, given her current state of calibration.

References

Audi, Robert (1998). *Epistemology: A Contemporary Introduction to the Theory of Knowledge.* Routledge.

BonJour, Laurence (1998). *In Defense of Pure Reason.* Cambridge University Press.

Carr, Jennifer (????).

Chalmers, David (2012). *Constructing the World.* Oxford University Press.

Chalmers, David J. (2011). "Revisability and Conceptual Change in "Two Dogmas of Empiricism"." *Journal of Philosophy*, 108(8): pp. 387–415.

Crimmins, Mark (2002). *Talk About Beliefs.* MIT Press.

Crimmins, Mark, and John Perry (1989). "The Prince and the Phone Booth: Reporting Puzzling Beliefs." *Journal of Philosophy*, 86(12): pp. 685–711.

Dretske, Fred (1981). *Knowledge and the Flow of Information.* MIT Press.

Dretske, Fred (1988). *Explaining Behavior: Reasons in a World of Causes.* MIT Press.

Evans, G. (1979). "Reference and Contingency." *The Monist*, 62(2): pp. 178–213.

Field, Hartry (2009). "Epistemology Without Metaphysics." *Philosophical Studies*, 143(2): pp. 249–290.

Frege, Gottlob (1948). "Sense and Reference." *Philosophical Review*, 57(3): pp. 209–230.

Frege, Gottlob (1956). "The Thought: A Logical Inquiry." *Mind*, 65(259): pp. 289–311.

Hawthorne, John (2002). "Deeply Contingent a Priori Knowledge." *Philosophy and Phenomenological Research*, 65(2): pp. 247–269.

Hedden, Brian (2013). "Options and Diachronic Tragedy." *Philosophy and Phenomenological Research*, 87(1).

Hintikka, Jaakko (1962). *Knowledge and Belief*. Ithaca, N.Y.,Cornell University Press.

Jackson, Frank (1982). "Epiphenomenal Qualia." *Philosophical Quarterly*, 32(April): pp. 127–136.

Kripke, Saul A. (1979). "A Puzzle About Belief." In *Meaning and Use*, Reidel, pp. 239–83.

Kripke, Saul A. (1980). *Naming and Necessity*. Harvard University Press.

Lewis, David (1982). "Logic for Equivocators." *Noûs*, 16(3): pp. 431–441.

Peacocke, Christopher (2004). *The Realm of Reason*. Oxford University Press.

Perry, John (1977). "Frege on Demonstratives." *Philosophical Review*, 86(4): pp. 474–497.

Quine, W. V. (1956). "Quantifiers and Propositional Attitudes." *Journal of Philosophy*, 53(5): pp. 177–187.

Quine, W. V. (1960). *Word and Object*. The Mit Press.

Quine, W. V. (1969). "Epistemology Naturalized." In *Ontological Relativity and Other Essays*, New York: Columbia University Press.

Quine, Willard V. O. (1953). "Two Dogmas of Empiricism." In *From a Logical Point of View*, New York: Harper Torchbooks, vol. 60, pp. 2–46.

Russell, Bertrand (1985). *The Philosophy of Logical Atomism*, vol. 29. Open Court.

Schiffer, Stephen (2004). "Skepticism and the Vagaries of Justified Belief." *Philosophical Studies*, 119(1-2): pp. 161–184.

Stalnaker, Robert (1984). *Inquiry*. Cambridge University Press.

Stalnaker, Robert (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press.

Stalnaker, Robert C. (1976). "Possible Worlds." *Noûs*, 10(1): pp. 65–75.

White, Roger (2006). "Problems for Dogmatism." *Philosophical Studies*, 131(3): pp. 525–57.

Williamson, Timothy (????). "How Deep is the Distinction Between A Priori and A Posteriori Knowledge?"

Williamson, Timothy (2007). *The Philosophy of Philosophy*. Blackwell Pub.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. New York,Macmillan.

Wright, Crispin (2002). "(Anti-)Sceptics Simple and Subtle: G. E. Moore and John McDowell." *Philosophy and Phenomenological Research*, 65(2): pp. 330–348.

Yablo, Stephen (2002). "Coulda, Woulda, Shoulda." In *Conceivability and Possibility*, Oxford University Press, pp. 441–492.