Lecture notes:
Graduate seminar on Conscious Life*, 2011*

BENJ HELLIE

April 7, 2011

Contents

Part I
Rationality

1 Rationality and psychology

What is the psychological?

We have various paradigms:

a. Deliberate bodily motions such as fingers moving to type various sentences;

b. Actions such as reasoning, conversing, playing, working;

c. Intentional states of belief, desire, hope, doubt, intention, (perhaps) knowledge;

d. Subjective qualitative states of sensing bodily and environmental conditions (seeing a red ball flying across a blue sky, having itch);

e. Emotions of anger and happiness;

f. Likes and dislikes;

g. Attentional states of looking at the red color of a ball or focusing on an itch;

h. ('Conscious') experiences.

A nice list of kinds of occurrence; what is it for something to 'have' (be in, undergo) one of them?

Let's sharpen the question.

- A *system* is a substantial particular in various intrinsic and extrinsic states, undergoing various intrinsic processes, engaging its environment in various extrinsic processes.

Suppose we are confronted with some system. Under what circumstances are these or those states or processes of the system of a psychological kind at all? And when one is, under what circumstances is it of this or that kind?

A popular answer to this question gives a significant role to 'constitutive rationality':

> Rational creatures are essentially agents. Representational mental states should be understood primarily in terms of the role that they play in the characterization and explanation of action. What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions. The agent has attitudes, pro and con, toward the different possible outcomes, and beliefs about the contribution which the the alternative actions would make to determining the outcome. One explains why an agent tends to act in the way he does in terms of such beliefs and attitudes. And, according to this picture, our conception of belief and of attitudes pro and con are conceptions of states which explain why a rational agent does what he does. [***stalnaker84, 4]

> Folk psychology says that a system of beliefs and desires tends to cause behavior that serves the subject's desires according to his beliefs. Folk psychology says that beliefs change constantly under the impact of perceptual evidence: we keep picking up new beliefs, mostly true, about our perceptual surroundings; whereupon our other beliefs (and our instrumental desires) change to cohere with these new beliefs. Folk psychology sets presumptive limits to what basic desires we can have or lack: *de gustibus non disputandum*, but still a bedrock craving for a saucer of mud would be unintelligible. Likewise it sets limits to our sense of plausibility: which hypotheses we find credible prior to evidence, hence which hypotheses are easily confirmed when their predictions come true. And it sets presumptive limits on what our contents of belief and desire can be. ... In short, folk psychology says that we make sense. It credits us with a modicum of rationality in our acting, believing, and desiring. [***lewis94, 320]

(The modern origin of this doctrine of constitutive rationality is found in this passage from *Word and Object*:

> This approach ill accords with a doctrine of 'prelogical mentality'. To take the extreme case, let us suppose that certain natives are said to accept as true certain sentences translatable in the form 'p and not p'. Now this claim is absurd under our semantic criteria. And, not to be dogmatic about them, what criteria might one prefer? Wanton translation can make natives sound as queer as one pleases. Better translation imposes our logic upon them, and would beg the question of prelogicality if there were a question to beg. [***quine60, 58]

This doctrine is also a big player in the philosophy of Davidson.)

I agree with these guys that rationality plays a big role in psychology but not with how they run with this idea. Here is how Lewis rolls it out.

Let's say that a story-frame is a set of sentences about a (possible) system: something of the form

σ is, at t, in states $s_1, s_2, \ldots$;

σ is, at t', in states $s'_1, s'_2, \ldots$;

$\ldots$;

σ is, over interval I, undergoing processes $p_1, p_2, \ldots$;

σ is, over interval I', undergoing processes $p'_1, p'_2, \ldots$;

$\ldots$;

$K_1(s_1); K_2(s_2); \ldots$;

$K'_1(s'_1); K'_2(s'_2); \ldots$;

$\ldots$;

$L_1(p_1); L_2(p_2); \ldots$;

$L'_1(p'_1); L'_2(p'_2); \ldots$;

$\ldots$;

s_{m_0} occurred because $p*^0_{n_1}, p*^0_{n_2}, \ldots, s*^0_{n_1}, s*^0_{n_2}, \ldots$ occurred;

$\ldots$;

p_{k_0} occurred because $p*^0_{l_1}, p*^0_{l_2}, \ldots, s*^0_{l_1}, s*^0_{l_2}, \ldots$ occurred;

$\ldots$;

Suppose we take a story-frame and make it specific by settling on a certain class of times and intervals, a certain number of states assigned to each of those times, and a certain class of processes assigned to each of the intervals; and by distributing a certain class of predicates over the states and processes. Then what we get is a *story*.

Sometimes, a story will use psychological predicates. And sometimes when we read such a story, it 'makes sense' as a story about someone's psychology: instead of saying things like 'one wanted to drink gasoline but one did not want to harm oneself or find out what it would be like or send a message or put on a show and one was not under orders and knew the likely effects of drinking gasoline' or 'one drank from the glass because one believed it contained gasoline and one wanted to drink a martini', it says thing like 'one drank from the glass because one believed

it contained a martini and one wanted to drink a martini'. Such a story makes the subject out as rational or reasonable. Say that a story that makes sense in this way has *psychological plausibility*.

Consider the class of stories with psychological plausibility: call it *folk psychology*. We can 'ramsify' folk psychology: going through the singular terms for psychological occurrences and psychological kind-predicates one by one, successively replace each with a fresh syntactically appropriate variable. When we do this, we get a big class of big complicated predicates containing only non-psychological vocabulary: logical vocabulary (including '='), 'because', and non-psychological predicates like 'puncture of the skin' and 'acoustic blast'.

Suppose we have some system which satisfies one of the (maximally specific) predicates Π of ramsified folk psychology. Then the substance of that system has psychology. Suppose also that occurrence o in that system of kind K is assigned to variable 'x', that Π contains the conjunct 'Fx', and that 'F' is the variable substituted for occurrences of 'is chopping garlic'. Then we say that o is a chopping of garlic and the substance of the system is chopping garlic throughout the duration of o.

According to Lewis, folk psychology is 'conceptually true': our practice of accepting its stories and rejecting the remaining stories involving psychological terms is based in our understanding of what those terms mean; where, moreover, the 'conceptual order' begins with the external world. He tells a 'myth of Rylean ancestors' who begin with a psychology-free language, and then introduce psychological expressions via implicit definition. This implicit definition consists in our asymmetric attitude toward the stories in and outside of folk psychology, and can be made explicit by taking ramsified folk psychology and issuing a proclamation like:

> Let *belief that roses are red*, *belief that violets are blue*, ...; refer to the satisfier of, respectively, the first predicate variable, the second predicate variable, ...; of ramsified folk psychology.

This is the sense in which psychology is, for Lewis, a sort of device for predicting and explaining behavior.

Lewis defends this claim on the grounds that it explains the 'odor of analyticity' around the 'platitudes of common-sense psychology' [***lewis72]—why it is that 'at some point ... weird tales of mental states that habitually offend against the principles of folk psychology stop making sense' [***lewis94].

2 Rationality and sense-making

Let us get further behind the notion of 'rationality' in play here. What is it for a story about someone to 'make sense'? The notion seems to have two components:

I. A synchronic notion of 'intelligibility': is the picture of the world one we can grasp? Are the goals ones that (perhaps stretching a bit) we can see something in?

II. A diachronic notion of 'reasonability': are the updates—adjustments of belief, commencements of action—fitting responses to the novel aspects of the situation in light of the prior picture and goals?

This section will say more about these notions: about what it is for someone's picture of the world to be intelligible, and about what it is for someone's update to be reasonable.

2.1 Intelligibility

Let's ignore intelligible desires and focus instead on intelligible pictures of the world. What is it for a story about someone's picture of the world (at a time) to make sense? If I am told that Sam believes that P, under what circumstances do I balk?

The answer here is straightforward. I find the claim that Sam believes that P unintelligible just if I find it unintelligible that P: just if I find the hypothesis that P unintelligible.

Here's an example. Consider the following evolving hypothesis about the world, to be understood as becoming more determinate monotonically:

i. Roses are red;

ii. Violets are blue;

iii. $2+2=4$;

iv. Grass is pink;

v. Bertrand Russell lived exactly 35,689 days;

vi. Benjamin Franklin was born in 1703;

vii. Right now, there are an even number of trees on Earth;

viii. It is not the case that violets are blue.

Here we are all right up through and including hypothesis (vii), even though our evolving hypothesis becomes more determinate, takes in the realm of the 'abstract' (iii), involves obvious truths (i, ii) and obvious falsehoods (iv), includes unobvious truths (v) and falsehoods (vi) and claims about which everyone is ignorant (vii). But when we get to (viii), we find that the hypothesis has now become *too* determinate to make any further sense of: that last stipulation cannot be combined with hypothesis (ii) into a coherent big hypothesis.

We find things to be exactly parallel when we consider a course of hypotheses not about the world but about Sam. Consider the following evolving hypothesis about Sam's beliefs at a time, understood as becoming more determinate monotonically:

i′. Sam believes that roses are red;

ii′. Sam believes that violets are blue;

iii′. Sam believes that $2+2=4$;

iv′. Sam believes that grass is pink;

v′. Sam believes that Bertrand Russell lived exactly 35,689 days;

vi′. Sam believes that Benjamin Franklin was born in 1703;

vii′. Sam believes that right now [at the time of attribution], there are an even number of trees on Earth;

viii′. Sam believes that it is not the case that violets are blue.

Once again, though Sam's hypothesized beliefs range over the 'abstract' (iii′), concern obvious truths (i′, ii′) and obvious falsehoods (iv′), includes unobvious truths (v′) and falsehoods (vi′) and claims about which everyone is ignorant (vii′), we encounter no difficulties in making sense of her. But when we get to (viii′), we find that the hypothesis about her belief has now become *too* determinate to make any further sense of: that last stipulation, if combined with hypothesis (ii′), makes an unintelligible big hypothesis about Sam's beliefs.

The following principle, therefore, seems very plausible:

> **Belief**
>
> The hypothesis that Sam believes that *P* renders Sam unintelligible just if the hypothesis that *P* is unintelligible.

Still, we should tread carefully:

- By saying that Sam is unintelligible I don't mean that she is *utterly* unintelligible: I mean something far weaker, like 'partly' unintelligible or 'less than fully' intelligible.

- Sam might be 'compartmentalized' or 'fragmented': bearing two competing belief states which are somewhat isolated from one another. David Lewis once realized that in his view, Nassau Street ran roughly (to within ten degrees) east–west, the train ran roughly north–south, and the two were roughly parallel. Collectively, these add up to an incoherent picture; separately, there is no incoherence; if Lewis is regarded as a composite of compartmentalized 'mini-agents' each of which is fully coherent, (Belief) can be preserved.

- We can distinguish 'positive' and 'negative' unintelligibility. Harman points out that, lacking the concept 'quantum number', one might fail to understand what is meant by the hypothesis that this electron's quantum number is five. If we are told, however, that Sam believes that this electron's quantum number is five, we do not thereby assess Sam as falling short of some psychological ideal. We can say then that our course of hypotheses (i)–(viii) is *positively* unintelligible while the hypothesis that this electron's quantum number is five is *negatively* unintelligible; the former is preserved under further information about lexical and compositional meaning while the latter is not.

- Graham Priest believes that some contradictions are true: let's suppose he accepts the English sentence 'violets are blue and it is not the case that violets are blue'. I can't make sense of his picture of the world while drawing it up with that sentence as I currently use it. But maybe I could 'think myself into' Priest's outlook: retrain myself into assigning a different inferential role to 'it is not the case that' than my ordinary role, an inferential role more like the one Priest assigns it. Perhaps I would undergo a sort of aspect shift through which I would come to find Priest's point of view intelligible. I'm somewhat inclined to want to assimilate my current attitude toward Priest to a case of negative unintelligibility.

There are nice questions about the range of intelligible situations. For example, assuming ordinary lexical and compositional meaning, are the following hypotheses intelligible or unintelligible?

- $7+5 \neq 12$;

- Hesperus is a planet but Phosphorus isn't;

- First-order Peano Arithmetic is complete;
- This apple is red all over and green all over;
- Horst is a kraut;
- $P \wedge \neg \Diamond P$;
- P at t; as of t', it had never been the case that P; $t < t'$;
- I don't exist;
- P and I don't believe that P;
- P and it might be that $\neg P$;
- P and I don't know whether P;
- Sam believes that P but assigns credence .001 to $\neg P$;
- Sam knows that P, but $\neg P$;
- Sam knows that P but I don't;
- I'm A-ing and I'm not trying to A;
- Sam was trying to write a book but had no idea she was trying to do so;
- Sam ran down the hill but she was never running down the hill;
- If P then Q, and P, but maybe not Q;
- P and there's some probability that $\neg P$.

I'm inclined to find these all unintelligible, though others no doubt differ.

2.2 Reasonability

Suppose that, at a certain time, Sam updates in a certain way: adjusts a belief, commences an action.

Two questions about this situation:

1. Was Sam reasonable in so updating?
2. Why did Sam so update?

These questions seem to be related. The answer to the former depends on the form of the class of answers to the latter. The update is reasonable just if a certain sort of explanation of Sam's update is available: namely, a 'rationalizing explanation'. But what is a rationalizing explanation?

Well first, what is an explanation? An answer—a true answer—to a 'why'-question, of course.

Here's how it works when the explanandum is the fact that Q. Suppose we accept P, and we accept that if P, normally Q. Then it follows that presumably, Q [***veltman96]. (This is non-monotonic, of course: if we go on to accept R and that if R, normally $\neg Q$, we withdraw the presumption that Q; and if we additionally go on to accept the exception to the initial rule to the effect that if $P \wedge R$, normally $\neg Q$, it follows that presumably $\neg Q$. But let's suppose it doesn't get that complicated.) Plausibly if Q, any batch of truths entailing presumably Q (perhaps or perhaps not also entailing, more strongly, Q) is an explanation of Q: explanations remove mystery, eliminate surprise; if some claim can be presumed given other claims we accept, it is no longer mysterious or surprising, but expected (given those claims). (It might not be an interesting helpful pleasant or nicely packaged explanation, but it would be an explanation nonetheless.) Often we observe conversations that go 'why P? because Q': here the response suppresses the conditional.

Second, what is a *rationalizing* explanation of an update? An explanation that 'makes sense' of the update, of course. Let's be a bit more concrete here:

> Suppose that we see Sam sorting a widget off to the right. Why did she do it? What is the rationalizing explanation for her action? What can make sense of it?

Here we need to hear from Sam. First a little background:

> My job is 'quality-controlling the widgets': when a widget comes down the conveyor belt, I am supposed to sort it off to the right just if it is defective, otherwise off to the left.
>
> And here I am in the factory, doing my job.

Now let's listen in on Sam's stream of consciousness over the interval within which she performs the act itself:

> Things are, going by looking, *thus*; so: this widget coming along is red! Red widgets are defective; so: this widget coming along is defective! I'm quality-controlling the widgets; so: sort this one off to the right!

We aren't supposing here that these words run through Sam's stream of consciousness. We are rather thinking of this narrative as a way into Sam's stream of consciousness: in imagining a situation by following this narrative, we imagine a stream of consciousness like Sam's. We might put this by saying that the 'content' of Sam's stream of consciousness is something like the 'content' of this narrative.

I find Sam's stream of consciousness to be entirely intelligible and reasonable. By contrast, the following streams of consciousness would not be reasonable:

> Things are, going by looking, *thus* [the same 'thus' as last time]; so: this widget coming along is **green!** ...
>
> Things are, going by looking, *thus*; so: this widget coming along is red! Red widgets are defective; so: this widget coming along is **not** defective! ...
>
> Things are, going by looking, *thus*; so: this widget coming along is red! Red widgets are defective; so: this widget coming along is defective! I'm quality-controlling the widgets; so: sort this one off to the **left!**

What is the ground of the contrast here?

Well, note that Sam's genuine stream of consciousness can be recast in reverse order as a course of explanations of her updates:

> Sort this widget coming along off to the right!
>
> But why do that? Because I'm quality-controlling the widgets, and the widget coming along is defective.
>
> But why believe the latter? Because red widgets are defective, and this widget coming along is red.
>
> But why believe the latter? Because things are, going by looking, *thus*.

These all strike me as good explanations. (At least they strike Sam that way.) By contrast, the following altered explanations are not good:

- Sort this widget coming along off to the **left!**

  But why do that? Because I'm quality-controlling the widgets, and the widget coming along is defective.

- The widget coming along is **not** defective!

  But why believe that? Because red widgets are defective, and this widget coming along is red.

- The widget coming along is **green!**

  But why believe that? Because things are, going by looking, *thus* [the same 'thus'].

It is because Sam's explanations are good and these others are bad, I suggest, that Sam's stream of consciousness is reasonable and the others are not.

Now note that the explananda in Sam's stream of consciousness are not facts but something more like commands addressed to oneself or commitments made to oneself: not 'the widget is red' or 'I believe that the widget is red' but 'believe that the widget is red!', where one is both master and slave, both binding oneself and being bound, both issuer and accepter of the command. How to square this sort of explanandum with our existing notion of explanation, which applies to factual explananda?

Let's try to make explanations of facts and explanations of commands as similar to one another as possible: any batch of truths and accepted conditional commands entailing *A*! or *believe that P*! explains it. So we can see Sam's course of explanations as running in something like the following way:

1. If things are, going by looking, *thus*, then: believe that the widget coming along is red!

   Things are, going by looking, *thus*;

   Ergo: believe that the widget coming along is red!

2. If red widgets are defective, and the widget coming along is red, then: believe that the widget coming along is defective!

   Red widgets are defective;

   The widget coming along is red;

   Ergo: believe that the widget coming along is defective!

3. When quality-controlling the widgets, if the widget coming along is defective, then: sort that widget off to the right!

   I'm quality-controlling the widgets;

   The widget coming along is defective;

   Ergo: sort that widget off to the right!

Each of these arguments is, intuitively, valid. By contrast, there seems to be no way to cast our intuitively unreasonable updates into the form of a valid explanation. Moreover, exclude any of the premisses from the stream of consciousness and

Sam's updates no longer look reasonable: for example, strike out the minor premiss of the first argument and Sam looks like an unreasonable 'clairvoyant'; strike out the action avowal in the third argument and Sam looks like an unreasonable 'automaton'.

Each of Sam's updates is reasonable because it is the concluding of a valid inference from premisses she accepts in the stream of consciousness. (Here we are thinking of 'concludings of inferences' as occurrences: plausibly, as 'achievements', occurrences that are essentially instantaneous. The notion of acceptance of a sentence 'in the stream of consciousness' is once again intended in the sense above, as something like 'entertaining the sentence is a way to think yourself into the stream of consciousness'.)

> **Update**
>
> The hypothesis that Sam updates by accepting (failing to accept) the command $A!$ against the background state of acceptances in the stream of consciousness of S renders Sam unintelligible just if the inference form $S \vdash A!$ is not valid (is valid). [***fix this]

We could use an addendum to our earlier discussion of synchronic intelligibility. After all, it is not so clear that our conditional commands—policies—count as things one could believe: accept, certainly, but they seem to have as little to do with capturing one's 'picture of the world' as do unconditional commands one accepts (which is not to deny that the latter can be the objects of belief, or that there may be beliefs that must track such acceptances—for instance, perhaps one cannot coherently accept $A!$ while failing to believe that one is A-ing).

As a substitute for a notion of an intelligible picture of the world, we might want a notion of a *sensible policy*. For instance:

- The policy in argument (1), perhaps, expands to the policy 'if Σ, and if 'Σ' and 'P' are equivalent (if one is unintelligible in diverging in one's attitudes toward them): then believe that $P!$ (here Σ and P are schematic letters over the 'sentences' of belief and perception, respectively); or more generally, a policy of updating belief to reflect the testimony of the senses;

- The policy in argument (2) looks like a special case of the policy 'if P, and Q, and $P; Q \vdash R$: then believe that $R!$'; or, more generally, a policy of updating belief to include all recognized consequences of recognized truths;

- The policy in argument (3) looks like a partial articulation of the policy commitments one takes on in the course of quality-controlling the widgets.

These all look pretty sensible: I could imagine accepting each of them as my own policy (we will have a great deal more to say about them as we go on); it is easy to conjure up alternatives that would be horrendous, by my lights.

So we can advance the following claim about what it is for an ascription of a policy to someone else to be intelligible:

> **Policy**
>
> The hypothesis that Sam has the policy Π renders Sam unintelligible just if the policy Π is not sensible.

Explanations have a generality to them: they remove mystery by bringing the particular under the general. Policies are similarly relatively general: they apply whenever the antecedent conditions are met, and when one is sensible, it is to be taken seriously by any rational agent.

2.3 Comments

1. We observe a significant 'transparency' in all three of our principles: rationality in belief with a certain content tracks intelligibility of the content itself; rationality in maintaining a certain policy tracks sensibility of the policy itself; rationality in updating via a certain inference form tracks validity of the inference form itself. This is quite striking for a number of reasons.

   (a) By saying 'this content is intelligible' or 'this policy is sensible' or 'this inference form is valid' I am of course expressing my own attitudes toward the entities in question. This transparency means that, to the extent that rationality is regarded by others as somehow regulative, everyone else with a psychology pretty much agrees with me in these attitudes. We find less similarity in other features of organisms.

   (b) A famous problem in 'modal epistemology' dates back to the central question motivating the *Tractatus*: why are the limits of the coherently believable the same as the limits of the possible? Why is it that one can coherently believe that P just if it is possible that P?

   We can now see this problem as an instance of a more general problem: for policies, the limits of the coherently followable are the same as the limits of the sensible; for updates, the limits of the rational track the limits of the valid inference forms. Piecemeal approaches, such as those which posit an a priori insight by which we survey an independently existing modal domain somewhat akin to the perceptual capacities by which we survey the world, are less satisfying because less unificatory.

Here is a sketch of a plausibly extensible answer to the *Tractatus* puzzle. The attitude of coherently entertaining the hypothesis that *P* is equivalent to the belief that *P* is possible. And since, as we shall see, the hypothesis that *P* is equivalent to the avowal of the belief that *P*, the attitude of coherently entertaining the hypothesis that *P* is equivalent to the attitude of coherently entertaining the hypothesis of avowal of the belief that *P* [***or maybe not]. This in turn is equivalent to the belief that avowal of the belief that *P* is possible. Assembling the strands, the belief that *P* is possible is equivalent to the belief that belief that *P* is possible; disquoting, *P* is possible just if belief that *P* is possible.

2. Commands have a multifarious visage. The command to believe that the widget coming along is red, issued in the concluding of argument (1), reappears as premiss in argument (2) as the proposition that the widget coming along is red.

   This too is striking. Why should it be so? And how can it be so?

   To see why it should be so, consider some alternatives. Suppose the conclusion of (1) were not an imperative but an indicative: 'the widget coming along is red' or 'I believe that the widget coming along is red'. In the former case the argument would not then be an explanation of an update but of a fact about the widget; and in the latter case, the concluding of the argument would not motivate. Alternatively, suppose the premiss of (2) were the imperative rather than the indicative; but conditionals test for *conditions*, which are propositions, expressed by indicatives. Alternatively, suppose the premiss of (2) were the indicative 'I believe that the widget coming along is red': but then the stream of consciousness would point back at itself rather than out at the world; and what's more, although this is grammatically in the indicative mood, semantically it is not an indicative but an 'evidential', and therefore does not express a proposition.

   How it *can* be so is for a notion of first-person equivalence to have rational regulative force. Suppose that self-issuance of a command requires its acceptance: one who self-commands *A*! is irrational unless one accepts the command *A*!. And suppose that acceptance of a command is equivalent to taking oneself to be following it—more explicitly, one accepts the command *believe that P!* just if one believes *I believe that P*; one accepts the command *cross the street!* just if one believes *I am crossing the street* (one does one but not the other on pain of irrationality). And suppose that *P* and *I believe that P* are equivalent (as per Moore's paradox).

Then the transition from the conclusion of (1) to the premiss of (2) unpacks as follows:

- One self-issues the command *believe that that widget is red!*;
- So one accepts the command *believe that that widget is red!*;
- So one believes *I believe that that widget is red*;
- So one believes *that widget is red*.

Moore's paradox is a long-standing philosophical mystery; this story (backed up by the Veltman-Gillies on the semantics of belief avowals) explains why it must arise.

We can imagine that something similar might arise in the case of actions. Presumably at some earlier point Sam issued the self-command *quality-control the widgets!*. This self-command shows up later as a premiss in argument (3). Action avowals, as we shall see, exhibit similar Moore-paradoxical phenomena, and will receive a similar Veltman-like semantic treatment.

If this is right, the stream of consciousness has what I think of as a sort of 'sawtooth' structure: policies link the ground floor level of facts about the world with the second floor level of commands to update; Moore-equivalence turns these commands back into ground floor facts.

3. As we shall see, Moore-paradoxical phenomena are best handled by denying that belief- and action-avowals express propositions. 'I believe that *P*' does not self-ascribe the property *belief that P*. But if not, then 'Sam believes that *P*' does not ascribe *belief that P* to Sam either. Similarly, 'I am *A*-ing' does not self-ascribe the property *being the agent of an A-ing in progress*. And if not, nor does 'Sam is *A*-ing' ascribe *being the agent of an A-ing in progress* to Sam.

4. These observations collectively make Lewis's 'myth of the Rylean ancestors' highly suspect:

   (a) Psychological avowals are given in the stream of consciousness, and do not need to be grounded in perceptual concepts of the nonpsychological;

   (b) Most psychological predicates do not express properties, and therefore are not fitting subjects for implicit definition via a primitively nonpsychological story;

(c) We have no idea which specific policies in regard to extracting evidence from sensory states might be reasonably followed without actually undergoing those states;

(d) Nor, since acts of *quality-controlling the widgets* are not ingredients of the world, could we have any idea of what the most effective way of performing one might be in the absence of a practical grasp of what it is to follow the self-command *quality-control the widgets!*;

(e) Moreover, the notion of the *concluding of an inference*, grasped straightforwardly on the basis of first-person resources, does not lend itself readily to analysis in objective terms (witness Davidsonian anxieties over 'causal deviance'); accordingly either the 'because' appearing in ramsified folk-psychology is a psychological notion or, if it is ramsified away, the result faces severe underdetermination threats;

(f) Transparency is explained rather trivially if I have my own stream of consciousness to go by as raw material in understanding others, a striking anomaly otherwise.

Familiarly also, (i) if third-person ascription goes by simulation, the 'odor' Lewis is smelling may be genuine but not that of analyticity; (ii) various third-personal theories are all subject to crushing objections of the 'underdetermination' variety; (iii) which suggests a meta-point, namely that our understanding of psychology does not consist in our acceptance of a theory; (iv) simulation is genuine and pervasive and fills in the gaps.

And yet at the same time, we will see that the proposition asserted by 'Sam believes that *P*' does turn out to be ultimately nonpsychological, and may in many cases be behavioral, thus 'paying our debt to behaviorism' [***lewis94].

3 Rationality more generally

Reasons for updating are parts of rational explanations of updating. People over in the ethics and moral psychology biz talk about other kinds of reasons as well: objective reasons, subjective reasons, yada yada. Here is how I see things in this regard.

3.1 Subjective reasons

Fred drank the liquid in that glass because he *believed* it was gin and tonic: too bad it was gasoline. What was Fred's reason for drinking the liquid? Here are some answers worth considering:

1. Just before drinking it, he would have answered 'it's gin and tonic'.

2. *We* wouldn't say 'it was gin and tonic'. Explanations have to be true.

3. We *might* say 'Fred *believed* it was gin and tonic'. That's true. But since if it is a psychological claim, it doesn't assert a proposition, it can't rationalize anything: rather, it either provides instructions for our extracting claim (1) or gives nonpsychological causal information.

I'm inclined to think the notion of a subjective reason doesn't make any sense. Or if it does, the claim that Fred's belief that *P* was a subjective reason for action amounts solely to the claim that here is how things were for Fred, subjectively: *P* is a reason for action.

3.2 Objective reasons

The liquid in this glass is gasoline, though Fred thinks it is gin and tonic. If I were Fred, I would not drink the liquid in that glass.

That case is pretty extraordinary: we might take away more representative lessons about practical reasoning from a more mundane case. Angela, engaging in a bit of urban tourism, is somewhat hungry. She finds herself in front of the local branch of Tasty Burrito, which dependably provides an OK lunch. What she doesn't know, but could—like I did—learn by combing through the *Downtown Weekly* she is holding, is that across the street lies Taco Bueno Sabor, where very delicious lunches are served. If I were Angela, I would cross the street and dine at Taco Bueno Sabor.

That's my story. Bill follows up with a different story:

> Something else Angela—like Hellie, it seems—does not know (but could—like I did—learn by asking the guys hanging around on the corner) is that up the block a bit lies Taco Caliente Peligroso, where mindblowingly delicious lunches are served. If *I* were Angela, I would go up the block and dine at Taco Caliente Peligroso.

I can put Bill's claim by saying that if *Bill* were Angela, Bill would go up the block and dine at Taco Caliente Peligroso.

Selena follows up with a different story still:

> Unlike Hellie and Bill, I prefer the familiar flavors of Tasty Burrito to the more rustic offerings available elsewhere on the block. So if I were Angela, I would stay right here and dine at Tasty Burrito.

If Selena were Angela, that is to say, Selena would dine right there at Tasty Burrito.

Now Mark follows up:

> I care nothing for the pleasures of the flesh: I care only about alleviating suffering. If I were Angela, I would purchase that potato and gnaw on it to alleviate hunger, and send the money I saved to Oxfam.

If Mark were Angela, Mark would forego restaurants entirely.

We can hear also from Ellen:

> If I were Angela, I would gnaw on that potato too, but not for Mark's reasons: he cares way too much about morality. Rather, because the scientific papers I have read prove that eating cooked food indoors shortens the lifespan by two decades. Well, I don't know. Angela might not be aware of these results or might be skeptical or might not care or might care but care more in the moment about the satisfactions of a hot meal. So I guess maybe if I were Angela, like Bill, I would go to Taco Caliente Peligroso. Well maybe I wouldn't. How would I know to do that? I'd have to ask those guys, and they might take that the wrong way. So maybe like Hellie I would go to Taco Bueno Sabor. But in order to do that I'd have to spend a bunch of time thumbing through *Downtown Weekly*, and I guess as Angela I'm hungry right now. So I guess I'd just go to Tasty Burrito. Or maybe I wouldn't even be facing this decision: rather than participating in urban tourism, I'd be in the office working on a paper. Well, maybe Mark is right: instead I'd be in the field nursing the wounded. But how would I get myself into a position of caring so much about morality? I'd have to spend a lot of time retraining myself into a different lifestyle and given the demands of the job that's time I don't have ...

Ellen tells a rather more complex story than any of our other conversants.

Plausibly, claims like 'if S were T, S would A' concern the doings of an agent with a stream of consciousness that is in some salient way a 'blend' of S's and T's: let $[S/T]_c$ be the agent blended in the way salient in context c. The logical form of the claim is something along the following lines: 'Let's hear from an agent whose stream of consciousness is the salient sort of blend of yours and mine: blah blah blah; so: A!'. (Here 'blah blah blah' narrates the salient blend.) To the extent that 'if I were you, I would A' serves to motivate the audience to A, it does so perhaps by motivating the audience to become the blended agent, who then issues the predicted self-command 'A!'.

In blending agents, what is preserved, and what is held fixed? As our examples show, there seems to be almost no discipline to this practice: we can add or subtract information, accommodate habits of theoretical reasoning or not,[1] accommodate the costs of acquiring information or not, flip around matters of personal taste or the relative weights of morality and self-concern, accommodate the costs of altering such matters of 'value' or not.

To the extent that notions like that of a 'reason' or a 'reasonable response' or what one 'ought' to do are understood along the lines suggested in the discussion of this section, as based in our notion of a rational update, a notion of what I 'objectively ought' to do could only be understood as what some Very Special Agent would do, if they were me. (When an ethical law of the form, 'Thou shalt ...' is laid down, one's first thought is, 'And what if I do not do it?'—*Tractatus* 6.422.) The apparent lack of discipline to our practice of blending agents suggests, to me at least, that there is probably no natural such Very Special Agent. The less liberally-minded, no doubt, will disagree.

[1] This factor is recognized in *The Lion in Winter*:

> Henry II: Good God, woman, face the facts.
>
> Eleanor of Aquitaine: Which ones? We have so many.

This concern seems to undermine the following case for the 'objectivity' of 'right' found in [***ross02]: what Bill ought to do is what he would do if he knew everything—which, if he does not, might be different from what in Bill's actual view he ought to do. I find it hard to make sense of the hypothesis that Bill 'knows everything'.

Part II
Test semantics and psychology

4 Basics

- A *language* is a class of expressions, usually generated recursively from an effectively specifiable *lexicon* or class of primitive expressions (some of which are designated as the 'logical' vocabulary) by a finite class of *formation rules* (some of the expressions of a language are designated as 'sentences').
- A *frame* is a 'tuple' (finite sequence), the elements of which are objects or sets of objects used in the specification of semantic properties.
- A *base valuation* function from the non-logical elements of the lexicon of a language to the entities contained within the model.
- A *model* is a frame concatenated with a base valuation function.

5 Propositional logic

5.1 A language

Consider the 'propositional' language L_0, specified in the following way:

Lexicon of L_0

- The non-logical vocabulary consists of an infinite set of 'proposition letters' $\Pi_0 = \{P_0, P_1, \ldots\}$;[2]
- The logical vocabulary consists of the *unary operator* $\neg$, the *binary operator* $\wedge$, and *left and right parentheses*.

Formation rules for L_0

- Every proposition letter is a sentence;
- If φ is a sentence, $\neg\varphi$ is a sentence;

[2]Notation: we are suppressing quotation marks because context explicitly distinguishes use and mention. It would be more strictly correct to say that the proposition letters are the objects in the set $\{\text{'}P_0\text{'}, \text{'}P_1\text{'}, \ldots\}$.

- If φ and ψ are sentences, $(\varphi \wedge \psi)$ is a sentence;[3]
- Nothing else is a sentence.

We will sometimes use the symbols $\vee$ and $\supset$ as informal abbreviations: $(\varphi \vee \psi) := \neg(\neg\varphi \wedge \neg\psi)$; $(\varphi \supset \psi) := (\neg\varphi \vee \psi)$.

The notion of 'scope' is useful. If a sentence is of form $\neg\varphi$, then the leftmost occurrence of negation 'takes scope over' every other operator in φ; a sentence is of form $(\varphi \wedge \psi)$, then the occurrence of conjunction associated with the outermost parentheses 'takes scope over' every other operator in either φ or ψ. For example, in the sentence $\neg(P_2 \wedge P_4)$, the negation takes scope over (or 'has wide scope with respect to') the conjunction; in the sentence $(\neg P_2 \wedge P_4)$, the conjunction takes scope over the negation (which, equivalently, 'has wide scope with respect to' the conjunction).

5.2 A frame

The frame $\mathscr{F}_0$[4] is:

- The one-element sequence $\langle W \rangle$; where
- W is the set $\{w_0, \ldots, w_2\}$; where
- The w_i are arbitrary objects, the exact nature of which is unimportant for our purposes—all that matters about them is that they be *numerically distinct*—but we will refer to them as 'possible worlds'.

5.3 A base valuation function

Let the function[5] $V : \Pi_0 \mapsto 2^W$ be such that:

[3]Notation: we use italic letters as the proposition letters of L_0 and lower-case greek letters as schematic sentence letters. We could put this rule more explicitly as follows: If φ and ψ are sentences, $\ulcorner(\varphi \wedge \psi)\urcorner$ is a sentence. Here the corner quotes are used to mark out the concatenation of a sequence of expressions, the members of which occur in the order in which they or their representations appear within the corner quotes. To be completely explicit, $\ulcorner(\varphi \wedge \psi)\urcorner$ is the expression '('$^\frown\varphi^\frown$'$\wedge$'$^\frown\psi^\frown$')'—namely, the string consisting of a left parenthesis concatenated with whatever expression 'φ' is contextually understood as standing in for, concatenated with the conjunction symbol, concatenated with whatever expression 'ψ' is contextually understood as standing in for, concatenated with a right parenthesis.

[4]Notation: we use cursive letters to represent *defined objects*, italic letters to represent *primitives*—the idea is that once the character of the primitives in a frame have been fixed, and some rules for getting from primitives to defined objects have been fixed, then so have the character of the defined objects. Frames themselves are defined objects, as are (as we will soon see) models.

[5]The specification placed on V immediately to follow means that the *domain* of V is a subset of Π_0 and the *range* of V is a subset of 2^W—more precisely, considering V as a *function* (a set of

- V is defined on $\{P_0, \ldots, P_7\}$:
  - $V(P_0) = \varnothing$;
  - $V(P_1) = \{w_0\}$;
  - $V(P_2) = \{w_1\}$;
  - $V(P_3) = \{w_2\}$;
  - $V(P_4) = \{w_0, w_1\}$;
  - $V(P_5) = \{w_0, w_2\}$;
  - $V(P_6) = \{w_1, w_2\}$;
  - $V(P_7) = \{w_0, w_1, w_2\}$;
- V is undefined otherwise.

5.4 A model

Let the model $\mathscr{M}_0 = \langle W, V \rangle$.

5.5 A semantic valuation function

The base valuation function V contained within our model $\mathscr{M}_0$ provides semantic values for the nonlogical elemental expressions of L_0. In order to assign semantic values to the complex sentences of L_0, we need to state some recursion rules. Later we will speak of something's holding in 'all models'—by which we will mean 'all models in which these (or the otherwise contextually appropriate) recursion rules are true'.

Recursion rules for classical connectives

Let $\|\cdot\|_{\mathscr{M}_0} : L_0 \mapsto 2^W$ have the following properties:

- If $\varphi \in \Pi_0, \|\varphi\|_{\mathscr{M}_0} = V(\varphi)$;
- If φ is of the form $\neg\psi$, $\|\varphi\|_{\mathscr{M}_0} = W \setminus \|\psi\|_{\mathscr{M}_0}$;
- If φ is of the form $(\psi \wedge \rho)$, $\|\varphi\|_{\mathscr{M}_0} = \|\psi\|_{\mathscr{M}_0} \cap \|\rho\|_{\mathscr{M}_0}$.

---

ordered pairs such that if $\langle a,b\rangle, \langle c,d\rangle \in V$ and $a = c$, then $b = d$), if $\langle a,b\rangle \in V$, then $a \in \Pi_0$ and $b \in 2^W$.

2^W is the 'power set' of W, or the set of all subsets of W.

Informally, what we are saying in placing this restriction on V is that V is to take a proposition letter and return the set of exactly those worlds at which that proposition letter is true.

Examples

- Consider $\neg(P_2 \wedge P_4)$: $\|\neg(P_2 \wedge P_4)\|_{\mathscr{M}_0} = W \setminus \|(P_2 \wedge P_4)\|_{\mathscr{M}_0}$; which in turn is equal to $W \setminus (\|P_2\|_{\mathscr{M}_0} \cap \|P_4\|_{\mathscr{M}_0})$; which in turn is equal to $W \setminus (V(P_2) \cap V(P_4))$; which in turn is equal to $W \setminus (\{w_1\} \cap \{w_0, w_1\})$; which in turn is equal to $W \setminus \{w_1\}$; which in turn is equal to $\{w_0, w_2\}$.

- Consider $(\neg P_2 \wedge P_4)$: $\|(\neg P_2 \wedge P_4)\|_{\mathscr{M}_0} = \|\neg P_2\|_{\mathscr{M}_0} \cap \|P_4\|_{\mathscr{M}_0}$; which in turn is equal to $(W \setminus \|P_2\|_{\mathscr{M}_0}) \cap \|P_4\|_{\mathscr{M}_0}$; which in turn is equal to $(W \setminus V(P_2)) \cap V(P_4)$; which in turn is equal to $(W \setminus \{w_1\}) \cap \{w_0, w_1\}$; which in turn is equal to $\{w_0, w_2\} \cap \{w_0, w_1\}$; which in turn is equal to $\{w_0\}$.

These examples show that 'scope' matters: the narrow-scope negation and wide-scope negation sentences have different semantic values.

6 Entailment

Notation: we will say that $\models_w^{\mathscr{M}} \varphi$ just if $w \in \|\varphi\|_{\mathscr{M}}$.

So in particular, if $\models_w^{\mathscr{M}_0} \neg(P_2 \wedge P_4)$, $w \in \{w_0, w_2\}$; while if $\models_w^{\mathscr{M}_0} (\neg P_2 \wedge P_4)$, $w \in \{w_0\}$. Accordingly, $\{w : \models_w^{\mathscr{M}_0} (\neg P_2 \wedge P_4)\} \subseteq \{w : \models_w^{\mathscr{M}_0} \neg(P_2 \wedge P_4)\}$.

That is noteworthy. Intuitively, the former claim is *stronger* than the latter claim. The latter says that at least one of P_2 and P_4 is false; the former is explicit in how this happens (P_2 is false but P_4 is not). So we should expect that the latter *follows from* or *is entailed by* or *is a consequence of* the former. Here we find that in model $\mathscr{M}_0$, the worlds at which the former is true is a (proper) subset of the worlds at which the latter is true. Can we use this relationship to uncover the nature of entailment?

Sort of, but we have to be careful. It is not enough to say that φ entails ψ just if in *some* model $\mathscr{M}$, $\{w : \models_w^{\mathscr{M}} \varphi\} \subseteq \{w : \models_w^{\mathscr{M}} \psi\}$. After all, note that $\{w : \models_w^{\mathscr{M}_0} P_2\} \subseteq \{w : \models_w^{\mathscr{M}_0} P_6\}$. But we do not want to say that P_2 entails P_6! These are arbitrary symbols with no intrinsic significance: their role in our enterprise is exhausted by their capacity to serve as things on which negation and conjunction can operate.

This undesirable prediction would be avoided if we instead said this:

- φ entails ψ just if in **every** model $\mathscr{M}$, $\{w : \models_w^{\mathscr{M}} \varphi\} \subseteq \{w : \models_w^{\mathscr{M}} \psi\}$.

The valuation function V is part of our model $\mathscr{M}_0$; relative to an alternative model $\mathscr{M}_0'$, another valuation V' could be employed such that $V(P_2) = V'(P_6)$ and $V(P_6) = V'(P_2)$. On this model, P_2's semantic value would *not* be a subset of P_6's.

Another way of putting the entailment condition is this:

- φ entails ψ just if in every model $\mathscr{M}$, $\|\varphi\|_{\mathscr{M}} \subseteq \|\psi\|_{\mathscr{M}}$.

We will sometimes say that φ is a *triviality* to mean that $\top$ entails φ; equivalently, in every model $\mathscr{M}$, $\|\varphi\|_{\mathscr{M}} = W$.

7 Modalities

We now turn to the properties of a more complex language capable of expressing our notions of necessity and possibility.

7.1 A language

Lexicon of L_1

- The lexicon of the language L_0; together with
- The logical 'necessity operator' $\Box$.

Intuitively, $\Box\varphi$ means 'it is necessarily the case that φ'.

Formation rules for L_1

- The formation rules of the language L_0; amended with the proviso that
- If φ is a sentence, $\Box\varphi$ is a sentence.

We also introduce a $\Diamond$ as an informal abbreviation using the following rule: $\Diamond\varphi := \neg\Box\neg\varphi$.

7.2 A primitive semantic valuation function

Recursion rules for the necessity operator

Let $\|\cdot\|_{\mathscr{M}_0} : L_1 \mapsto 2^W$ have the following properties:

- All those properties already stipulated to hold of the semantic valuation function relative to $\mathscr{M}_0$ with domain L_0; together with
- If φ is of the form $\Box\psi$, $\|\varphi\|_{\mathscr{M}_0} = \{w \in W : (\forall w' \in W)(w' \in \|\psi\|_{\mathscr{M}_0})\}$.

Examples

- $\|\Box P_7\|_{\mathscr{M}_0} = \{w \in W : (\forall w' \in W)(w' \in \|P_7\|_{\mathscr{M}_0})\}$; this in turn is equal to $\{w \in W : (\forall w' \in W)(w' \in \{w_0, w_1, w_2\})\}$; this in turn is equal to $\{w \in W : \top\}$;[6] this in turn is equal to W.

- $\|\Box P_6\|_{\mathscr{M}_0} = \{w \in W : (\forall w' \in W)(w' \in \|P_6\|_{\mathscr{M}_0})\}$; this in turn is equal to $\{w \in W : (\forall w' \in W)(w' \in \{w_1, w_2\})\}$; this in turn is equal to $\{w \in W : \bot\}$; this in turn is equal to $\varnothing$.

- $\|\neg\Box P_6\|_{\mathscr{M}_0} = W \setminus \{w \in W : (\forall w' \in W)(w' \in \|P_6\|_{\mathscr{M}_0})\}$; since as we have seen the set being subtracted from W is equal to $\varnothing$, $\|\neg\Box P_6\|_{\mathscr{M}_0} = W$.

- $\|\Box\neg P_0\|_{\mathscr{M}_0} = \{w \in W : (\forall w' \in W)(w' \in \|\neg P_0\|_{\mathscr{M}_0})\}$; this in turn is equal to $\{w \in W : (\forall w' \in W)(w' \in (W \setminus \|P_0\|_{\mathscr{M}_0}))\}$; this in turn is equal to $\{w \in W : (\forall w' \in W)(w' \in (W \setminus \varnothing))\}$; this in turn is equal to $\{w \in W : (\forall w' \in W)(w' \in W)\}$; this in turn is equal to $\{w \in W : \top\}$; this in turn is equal to W.

- $\|\Box\Box P_7\|_{\mathscr{M}_0} = \{w \in W : (\forall w' \in W)(w' \in \|\Box P_7\|_{\mathscr{M}_0})\}$; as we have seen, every world in W is in $\|\Box P_7\|_{\mathscr{M}_0}$, so the condition is trivially met; so $\|\Box\Box P_7\|_{\mathscr{M}_0} = W$.

Exercise

Determine $\|\Box\neg\Box P_6\|_{\mathscr{M}_0}$.

8 Frames with accessibility

We have just vaguely gestured at the point that on our current system, $\Box\varphi$ entails $\Box\Box\varphi$ and $\neg\Box\varphi$ entails $\Box\neg\Box\varphi$—what is necessarily true is necessarily so; and the same goes for what isn't.

This system is not very interesting. Preserving the current definition of a frame and the current formation rule for $\Box$, we get that $\models^{\mathscr{M}}_{w} \Box\varphi$ just if $\models^{\mathscr{M}}_{w'} \Box\varphi$. Facts about necessity aren't contingent. We can make them contingent by adding a new kind of ingredient to our frames and complexifying the influence of $\Box$ on semantic value accordingly.

[6] Notation: we let '$\top$' stand for some trivially true condition, '$\bot$' stand for some trivially false condition.

8.1 A frame

Let the frame $\mathscr{F}_1$ be:

- $= \langle W, R \rangle$, where $R : W \mapsto W$.

R is known as an 'accessibility' relation: informally, if $R(w) = w'$, we say that w 'sees' w'. The thought is that w' is a possibility not absolutely, but on the assumption that w is actual. Instead of $R(w) = w'$, we will write Rww' or wRw'.

8.2 A model

The model $\mathscr{M}_1$ is:

- $\mathscr{F}_1$ concatenated with the valuation function V.

8.3 A semantic valuation function

I haven't told you all about $\mathscr{M}_1$ yet because I haven't said anything about R, but we have enough to provide a semantic valuation clause for $\Box$. We replace the clause from 4.2 with:

- $\|\Box\varphi\|_{\mathscr{M}_1} = \{w : (\forall w' : wRw')(\models^{\mathscr{M}_1}_{w'} \varphi)\}$;
  equivalently, $\|\Box\varphi\|_{\mathscr{M}_1} = \{w : \{w' : wRw'\} \subseteq \|\varphi\|_{\mathscr{M}_1}\}$;
  equivalently, $\|\Box\varphi\|_{\mathscr{M}_1} = \{w : \{w' : wRw'\} \cap \|\varphi\|_{\mathscr{M}_1} = \{w' : wRw'\}\}$.

8.4 Properties of frames

Some important conditions on R correspond to certain axioms governing the box:

- R is *reflexive* (for all w, wRw) just if $\|\Box\varphi \supset \varphi\|_{\mathscr{M}} = W$ for every $\mathscr{M}$ in which R appears;
- R is *transitive* (for all w, w', w'': if wRw' and $w'Rw''$, then $w'Rw''$) just if $\|\Box\varphi \supset \Box\Box\varphi\|_{\mathscr{M}} = W$ for every $\mathscr{M}$ in which R appears;
- R is *symmetric* (for all w, w': if wRw' then $w'Rw$) just if $\|\neg\Box\varphi \supset \Box\neg\Box\varphi\|_{\mathscr{M}} = W$ for every $\mathscr{M}$ in which R appears.

Exercise

- Restate these conditions using diamonds instead of negations.

8.5 Systems of modality

There are weird labels for the axiom schemata on the box. The first of these is called T; the second is called $\mathsf{4}$; and the third is called $\mathsf{5}$. (There are a lot of other axiom schemata that have been given similar weird names.)

We speak of 'systems' of modal logic:

- The system **K** has as axioms every tautology of L_0, T, the distribution axiom schema $\Box(\varphi \supset \psi) \supset (\Box\varphi \supset \Box\psi)$; and as a rule of inference the 'necessitation' rule (if φ is a theorem, $\Box\varphi$ is a theorem).
- The system **S4** is $\mathbf{K} + \mathsf{4}$; the system **S5** is $\mathbf{S4} + \mathsf{5}$.
- In the system **S5**, on which R is an *equivalence relation*, the behavior of the box is equivalent to that displayed using the valuation rule from $\mathscr{M}_0$.

9 Introducing contexts

Suppose we wish to speak of what is actually the case. We want to say that, while it is necessary that what is so is so, it is merely contingent that what is *actually* so is so.

9.1 A language

To get L_2, we add an operator A to the logical lexicon of L_1, along with the following formation rule:

- If φ is a sentence, $\mathsf{A}\varphi$ is a sentence.

Our intuition can be then cast into a formalism as follows: while $\Box(\varphi \supset \varphi)$ is a triviality, $\Box(\varphi \supset \mathsf{A}\varphi)$ is not.

9.2 A frame and a model

Let the frame $\mathscr{F}_2$ be:

- The triple $\langle W, R, \mathscr{C} \rangle$, where
- $\mathscr{C} = W$.

And let the model $\mathscr{M}_2$ be:

- $\mathscr{F}_2$ concatenated with the valuation function V.

9.3 A semantic valuation function

The added complexity of our frame is to allow us to have two independently varying parameters built into semantic values: one which keeps track of which world is the *object* of discussion; and one which keeps track of which world is the *context* of discussion. So even in speaking of what is necessarily the case, in such a way that every world comes in as an object of discussion, we maintain a grounding in the world which is the context of discussion. This allows us to treat A as 'reaching back' to the context of discussion in a way that is not influenced by the machinations of $\Box$. The members of W will play the former role of providing a parameter dedicated to $\Box$; the members of $\mathscr{C}$ will play the latter role of providing a parameter dedicated to A.

Recursion rules for a modal language with A

Let $\|\cdot\|_{\mathscr{M}_2} : L_2 \mapsto 2^{W\times\mathscr{C}}$[7] work as follows:

- If $\varphi \in \Pi_0, \|\varphi\|_{\mathscr{M}_2} = V(\varphi)\times\mathscr{C}$;[8]
- If φ is of the form $\neg\psi$, $\|\varphi\|_{\mathscr{M}_2} = (W\times\mathscr{C}) \setminus \|\psi\|_{\mathscr{M}_2}$;
- If φ is of the form $(\psi\wedge\rho)$, $\|\varphi\|_{\mathscr{M}_2} = \|\psi\|_{\mathscr{M}_2} \cap \|\rho\|_{\mathscr{M}_2}$;
- If φ is of the form $\Box\psi$, $\|\varphi\|_{\mathscr{M}_2} = \{\langle w,c\rangle : \{\langle w',c\rangle : wRw'\} \subseteq \|\psi\|_{\mathscr{M}_2}\}$;
- If φ is of the form $\mathsf{A}\psi$, $\|\varphi\|_{\mathscr{M}_2} = \{\langle w,c\rangle : \langle c,c\rangle \in \|\psi\|_{\mathscr{M}_2}\}$.

Examples

Let's verify that this does what we want it to:

- $\|\Box(\varphi \supset \varphi)\|_{\mathscr{M}} = \{\langle w,c\rangle : \{\langle w',c\rangle : wRw'\} \subseteq \|\varphi\supset\varphi\|_{\mathscr{M}}\}$. Now, $\|\varphi\supset\varphi\|_{\mathscr{M}} = \|\neg\varphi\|_{\mathscr{M}} \cup \|\varphi\|_{\mathscr{M}} = (W\times\mathscr{C}\setminus\|\varphi\|_{\mathscr{M}}) \cup \|\varphi\|_{\mathscr{M}} = W\times\mathscr{C}$.
- By contrast, $\|\Box(\varphi\supset\mathsf{A}\varphi)\|_{\mathscr{M}} = \{\langle w,c\rangle : \{\langle w',c\rangle : wRw'\} \subseteq \|\varphi\supset\mathsf{A}\varphi\|_{\mathscr{M}}\}$. Now $\|\varphi\supset\mathsf{A}\varphi\|_{\mathscr{M}} = \|\neg\varphi\|_{\mathscr{M}} \cup \|\varphi\|_{\mathscr{M}} = (W\times\mathscr{C}\setminus\|\varphi\|_{\mathscr{M}}) \cup \|\mathsf{A}\varphi\|_{\mathscr{M}}$.

  To evaluate this latter formula, set $\mathscr{M} = \mathscr{M}_2$; set $\varphi = P_4$; and focus on the world-context pair $\langle w,c\rangle = \langle w_0, w_2\rangle$.

[7] $W\times\mathscr{C}$ is the set of all ordered pairs $\langle w,c\rangle$ such that $w\in W$ and $c\in\mathscr{C}$. So $2^{W\times\mathscr{C}}$ is the power set of this set: the set containing each of its subsets.

[8] By this I mean that if w is among the worlds assigned to $\varphi\in\Pi_0$ by V, then for every context $c\in\mathscr{C}$, $\langle w,c\rangle \in \|\varphi\|_{\mathscr{M}_2}$. Which context you speak from doesn't influence which base-line proposition you utter with an elementary sentence.

Note that $\langle w_0, w_2\rangle \in \|P_4\|_{\mathscr{M}_2}$ (because $w_0 \in V(P_4)$ and for all w, c, if $w \in V(P_4)$, $\langle w, c\rangle \in \|P_4\|_{\mathscr{M}_2}$).

And note that $\langle w_0, w_2\rangle \notin \|\mathsf{A}P_4\|_{\mathscr{M}_2}$ (because $\|\mathsf{A}P_4\|_{\mathscr{M}_2} = \{\langle w, c\rangle : \langle c, c\rangle \in \|P_4\|_{\mathscr{M}_2}\}$; and $\langle w_2, w_2\rangle \notin \|P_4\|_{\mathscr{M}_2}\}$ because $w_2 \notin V(P_4)$).

So we notice that $\langle w_0, w_2\rangle$ is not in $W \times \mathscr{C} \setminus \|P_4\|_{\mathscr{M}_2}$; and it is also not in $\|\mathsf{A}P_4\|_{\mathscr{M}_2}$; so it is not in their union. So $\|\Box(P_4 \supset \mathsf{A}P_4)\|_{\mathscr{M}_2} \neq W \times \mathscr{C}$; so it is not a triviality: as desired.

9.4 Philosophical excursus on 'meaning'

Can we still think of 'meanings' as sets of worlds? Or do we need to think of them as sets of pairs of contexts and worlds?

The question arises because on the one hand, P_4 and $\mathsf{A}P_4$ have different semantic values: are true at different pairs of contexts and worlds.

But, on the other hand, we can also construct a sense in which P_4 and $\mathsf{A}P_4$ are true 'at the same worlds':

$$\|P_4\| = \{\langle w_0, w_0\rangle, \langle w_0, w_1\rangle, \langle w_0, w_2\rangle, \langle w_1, w_0\rangle, \langle w_1, w_1\rangle, \langle w_1, w_2\rangle\},$$

while

$$\|\mathsf{A}P_4\| = \{\langle w_0, w_0\rangle, \langle w_1, w_0\rangle, \langle w_2, w_0\rangle, \langle w_0, w_1\rangle, \langle w_1, w_1\rangle, \langle w_2, w_1\rangle\};$$

so think of the 'worlds at which φ is true' as $\{w : \langle w, w\rangle \in \|\varphi\|\}$—the so-called 'diagonalization' of $\|\varphi\|$—intuitively, as something like the worlds within which φ is correctly affirmed.

So: do P_4 and $\mathrm{A}P_4$ differ in meaning?

On the one hand, they appear in sentences with divergent truth-conditions (which is why we assigned them divergent semantic values).

On the other hand, to affirm one and to affirm the other are in a sense 'one and the same': we see no difference between how affirming one says things are and how affirming the other says they are.

I'm somewhat inclined to suspect this shows that we should distinguish two kinds of meaning: *logical meaning*, which is sensitive to the distinction between the sentences; and *affirmational meaning*, which isn't.

This distinction is in on the ground floor of analytic philosophy, being recognized by Frege in his distinction between the 'content stroke' and the 'judgement stroke'. There are other reasons for endorsing it as well: $\neg P$ has some overlap in meaning with P, but in the latter one does while in the former one emphatically does not affirm that P.

The phenomena are related like this: affirmational meaning is that to which we have direct phenomenological access in terms of the 'picture of the world' we take to be encoded in an affirmed sentence (in affirming P, one pictures the world as one in which snow is white); logical meaning is more of a theoretical posit, used to explain patterns in affirmational meaning (the logical meaning of P, constant between $\neg P$ and P simpliciter relates the two by revealing that their affirmational meanings exclude one another, or some such).

In our formal framework, we might then identify logical meaning with semantic value, and affirmational meaning with the diagonalization of semantic value.

Since entailment is a relation between sentences 'taken whole' rather than between sentences 'as potential', we would want to introduce a second notion of entailment. Letting

- $\Delta_{\mathscr{M}}\varphi = \{w : \langle w,w\rangle \in \|\varphi\|_{\mathscr{M}}\}$,

We can define the following notion of entailment:

- φ Δ-*entails* ψ just if in every model $\mathscr{M}$, $\Delta_{\mathscr{M}}\varphi \subseteq \Delta_{\mathscr{M}}\psi$.

This would have the result that in all cases, φ and $\mathsf{A}\varphi$ Δ-entail one another. We think of Δ-entailment, therefore, as a better candidate than classical entailment for providing the structure of affirmational meaning.

Its diagonalization is one object that can be extracted from a semantic value. Another is the *classical intension relative to a context*: intuitively, the class of worlds at which the expression with that semantic value is true if produced in that context. Let

- $I^c_{\mathscr{M}}\varphi = \{w : \langle w,c\rangle \in \|\varphi\|_{\mathscr{M}}\}$.

We can then define the following notions of entailment:

- φ *classically entails* ψ *relative to* c just if in every model $\mathscr{M}$, $I^c_{\mathscr{M}}\varphi \subseteq I^c_{\mathscr{M}}\psi$;
- φ *absolutely classically entails* ψ just if in every model $\mathscr{M}$, $\forall c$:

$$I^c_{\mathscr{M}}\varphi \subseteq I^c_{\mathscr{M}}\psi.$$

10 Epistemic modals

We note that 'certainly P' and 'P' are equivalent: one rightly accepts one just if one rightly accepts the other. And yet neither *entails* the other: taking the external

perspective, one might be certain of what is false. Similarly, 'it might be the case that P' and '$\neg P$' are incompatible, though neither entails the negation of the other. These facts are independent of whether by 'entails' we intend 'classical' or 'Δ'-type entailment. We need to construct the notion of equivalence in play here, which we can think of as *presupposition*.

10.1 A language

L_3 expands L_2 to include the operator B, along with the following formation rule:

- If φ is a sentence, $\mathsf{B}\varphi$ is a sentence.

The operator M is the dual of B, with $\mathsf{M}\varphi$ serving to abbreviate $\neg\mathsf{B}\neg\varphi$.

10.2 A frame and a model

Let $\mathscr{F}_3$ be:

- The triple $\langle W, R, \mathscr{C}\rangle$, where
- $\mathscr{C} = W \times 2^W$.[9]

Our image of context has resolved further detail: we think of a context as a pair of a world and a proposition. We will think of a particular context $c \in \mathscr{C}$ as an ordered pair $\langle w_c, i_c\rangle$, where intuitively w_c is the world in which c is located and i_c is a proposition representing the total body of information taken for granted in c.

Then let the model $\mathscr{M}_3$ be:

- $\mathscr{F}_3$ concatenated with the valuation function V.

10.3 A semantic valuation function

Recursion rules for a modal language with A and B

Let $\|\cdot\|_{\mathscr{M}_3} : L_3 \mapsto 2^{W\times\mathscr{C}}$ work as follows:

- If $\varphi \in \Pi_0, \|\varphi\|_{\mathscr{M}_3} = V(\varphi) \times \mathscr{C}$;
- If φ is of the form $\neg\psi$, $\|\varphi\|_{\mathscr{M}_3} = (W \times \mathscr{C}) \setminus \|\psi\|_{\mathscr{M}_3}$;
- If φ is of the form $(\psi \wedge \rho)$, $\|\varphi\|_{\mathscr{M}_3} = \|\psi\|_{\mathscr{M}_3} \cap \|\rho\|_{\mathscr{M}_3}$;

[9] Namely, the set consisting of every pair the first member of which is a world and the second member of which is a proposition.

- If φ is of the form $\Box\psi$, $\|\varphi\|_{\mathscr{M}_3} = \{\langle w,c\rangle : \{\langle w',c\rangle : wRw'\} \subseteq \|\psi\|_{\mathscr{M}_3}\}$;
- If φ is of the form $\mathsf{A}\psi$, $\|\varphi\|_{\mathscr{M}_3} = \{\langle w,\langle w_c,i_c\rangle\rangle : \langle w_c,\langle w_c,i_c\rangle\rangle \in \|\psi\|_{\mathscr{M}_3}\}$;
- If φ is of the form $\mathsf{B}\psi$, $\|\varphi\|_{\mathscr{M}_3} =$

$$\{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \|\psi\|_{\mathscr{M}_3})\}.$$

  Note that according to this this clause, the information taken for granted in a context resembles the diagonalization of semantic value more closely than it resembles the classical intension (relative to any context world).

Examples

- $\|P_4\|_{\mathscr{M}_3} = \{w_0,w_1\} \times \mathscr{C}$.
- $\|\mathsf{B}P_4\|_{\mathscr{M}_3} = \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \|P_4\|_{\mathscr{M}_3}\}$

$$= \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \{w_0,w_1\} \times \mathscr{C})\}$$

$$= W \times (W \times 2^{\{w_0,w_1\}}).$$

- $\|\neg\mathsf{B}P_4\|_{\mathscr{M}_3} = (W \times \mathscr{C}) \setminus \|\mathsf{B}P_4\|_{\mathscr{M}_3}$

$$= W \times (W \times (\{\varnothing\} \cup \{\mathbf{P} \in 2^W : w_2 \in \mathbf{P}\})).$$

- $\|\mathsf{B}\mathsf{B}P_4\|_{\mathscr{M}_3} = \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \|\mathsf{B}P_4\|_{\mathscr{M}_3}\}$

$$= \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in W \times (W \times 2^{\{w_0,w_1\}})\}$$

$$= W \times (W \times 2^{\{w_0,w_1\}}).$$

- $\|\mathsf{B}\mathsf{A}P_4\|_{\mathscr{M}_3} = \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \|\mathsf{A}P_4\|_{\mathscr{M}_3}\}$

$$= \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \{\langle w,\langle w_c,i_c\rangle\rangle : \langle w_c,\langle w_c,i_c\rangle\rangle \in \|P_4\|_{\mathscr{M}_3}\})\}$$

$$= \{\langle w,\langle w_c,i_c\rangle\rangle : (\forall w' \in i_c)(\langle w',\langle w',i_c\rangle\rangle \in \|P_4\|_{\mathscr{M}_3})\}$$

$$= W \times (W \times 2^{\{w_0,w_1\}}).$$

Exercises

Determine, relative to $\mathscr{M}$, the semantic values of the following sentences of L_3:

- $\mathsf{B}\neg\varphi$;
- $\mathsf{B}\wedge\varphi$;
- $\mathsf{B}\Box\varphi$.

What undesirable prediction would the following clause make regarding the semantic value of $\mathsf{BA}\varphi$?

- If φ is of the form $\mathsf{B}\psi$, $\|\varphi\|_{\mathscr{M}_3} =$

$$\{\langle w, \langle w_c, i_c\rangle\rangle : (\forall w' \in i_c)(\langle w', \langle w_c, i_c\rangle\rangle \in \|\psi\|_{\mathscr{M}_3})\}.$$

10.4 Presupposition

It is evident that neither of φ nor $\mathsf{B}\varphi$ will in general classically-entail the other; nor will either Δ-entail the other (assuming the obvious extension of that notion to cover the formulae of L_3). We need to explain their presuppositional equivalence in some other way.

Exercise

- Show that in general, $\mathsf{B}\varphi$ and $\mathsf{BB}\varphi$ classically entail one another.

The joint entailment of these claims suggests a route to the desired explanation. Roughly, the idea is that φ presupposes ψ just if $\mathsf{B}\varphi$ classically entails $\mathsf{B}\psi$: just if a context encoding the information in φ but failing to encode the information in ψ is metaphysically impossible.

We can formalize this idea as follows.

Let $[\![\cdot]\!]_{\mathscr{M}_3} : L_3 \mapsto \mathscr{C}$ be such that:

- $[\![\varphi]\!]_{\mathscr{M}_3} = \{i_c : (\forall w \in i_c)(\langle w, \langle w, i_c\rangle\rangle \in \|\varphi\|_{\mathscr{M}_3})\}$.

If we label $[\![\varphi]\!]_{\mathscr{M}_3}$ the *bodies of information supporting* φ (relative to $\mathscr{M}_3$), we can put this relationship as follows: the bodies of information supporting φ are just those present in the semantic value of $\mathsf{B}\varphi$.

Exercise

Show that, for all models $\mathscr{M}$:

- $[\![\neg\varphi]\!]_{\mathscr{M}} = 2^W \setminus [\![\varphi]\!]_{\mathscr{M}}$;
- $[\![\varphi \wedge \psi]\!]_{\mathscr{M}} = [\![\varphi]\!]_{\mathscr{M}} \cap [\![\psi]\!]_{\mathscr{M}}$;
- $[\![\mathsf{B}\varphi]\!]_{\mathscr{M}} = [\![\varphi]\!]_{\mathscr{M}}$;
- $[\![\mathsf{A}\varphi]\!]_{\mathscr{M}} = [\![\varphi]\!]_{\mathscr{M}}$.

With this apparatus of support in hand, we can define a notion of presupposition:

- φ presupposes ψ just if, for every model $\mathscr{M}$, $[\![\varphi]\!]_{\mathscr{M}} \subseteq [\![\psi]\!]_{\mathscr{M}}$.

Or, unpacking the definition of $[\![\cdot]\!]_{\mathscr{M}}$:

- φ presupposes ψ just if, for every model $\mathscr{M}$:

$$\{i_c : (\forall w \in i_c)(\langle w, \langle w, i_c\rangle\rangle \in \|\varphi\|_{\mathscr{M}})\}$$

$$\subseteq \{i_c : (\forall w \in i_c)(\langle w, \langle w, i_c\rangle\rangle \in \|\psi\|_{\mathscr{M}})\}.$$

10.5 Meaning

Can we locate a notion of the proposition expressed by a sentence of L_3—understood as something like a set of worlds the sentence represents one as located in, and which is adequate to our notion of presupposition? At present we don't have that, because presupposition has been cast as resting on relations between *sets* of bodies of information rather than between just bodies of information. We want to somehow 'project' these sets of bodies downward onto the bodies themselves. But the difficulty is doing this without losing any important structure gained by the added complexity.

To see what I mean, suppose we take the crude but simple approach of just smooshing the relevant bodies of information into a single body of information, as follows:

- $\lceil\varphi\rceil_{\mathscr{M}} = \{w : (\exists i_c \in [\![\varphi]\!]_{\mathscr{M}})(w \in i_c)\}$.

The information conveyed by φ—the proposition φ expresses—is the *smallest* amount of information common to all the contexts supporting φ. Assume this 'smooshing' conception for reductio.

Now, it seems to be a desirable feature of any notion of proposition that, if **P** is the proposition expressed by φ, the proposition expressed by $\neg\varphi$ is $W \setminus \mathbf{P}$: negation is a polarity-flipper with respect to propositions. So let us assume this principle.

So consider $\mathsf{B}P_4$. According to the smooshing conception, this expresses the proposition $\{w_0, w_1\}$ (exercise: show this). And consider $\neg\mathsf{B}P_4$. One way to fail to believe P_4 is to be totally uncertain; more formally, $W \in [\![\neg\mathsf{B}P_4]\!]$. So according to the smooshing conception, this expresses the proposition W. But of course $W \setminus \{w_0, w_1\} \neq W$. So the smooshing conception is at odds with our principle about negation, and should be rejected.

Let us try the only obvious alternative approach, of rendering the notion of the proposition expressed by a sentence always *relative to a body of information*. We implement the idea as follows:

- $\lceil\varphi\rceil^{i_c}_{\mathscr{M}} = \{w : \langle w, \langle w, i_c\rangle\rangle \in \|\varphi\|_{\mathscr{M}}\}$.

The content of the proposal is best brought out by looking at some examples.

Examples

- $\lceil P_4\rceil^{i_c}_{\mathscr{M}_3} = \{w : \langle w, \langle w, i_c\rangle\rangle \in \|P_4\|_{\mathscr{M}_3}\}$;
  Since the semantic value of P_4 is insensitive to the information coordinate, then no matter what i_c is, this set is equal to the diagonalization of that semantic value: namely, $\{w_0, w_1\}$.
- The same holds for $\mathsf{A}P_4$.
- $\lceil\mathsf{B}P_4\rceil^{i_c}_{\mathscr{M}_3} = \{w : \langle w, \langle w, i_c\rangle\rangle \in \|\mathsf{B}P_4\|_{\mathscr{M}_3}\}$.
  Now, $\|\mathsf{B}P_4\|_{\mathscr{M}_3}$, by contrast, is sensitive to the information coordinate but insensitive to either world coordinate. Accordingly, we should expect the proposition expressed by this sentence relative to a body of information to be world-insensitive or 'noncontingent': either necessary or impossible; but which of these noncontingent propositions is expressed, we should anticipate, will be sensitive to the character of i_c, as follows:
  - $\lceil\mathsf{B}P_4\rceil^{W}_{\mathscr{M}_3} = \{w : \langle w, \langle w, W\rangle\rangle \in \|\mathsf{B}P_4\|_{\mathscr{M}_3}\}$
    $= \{w : \langle w, \langle w, W\rangle\rangle \in W \times (W \times 2^{\{w_0, w_1\}})\} = \varnothing$;

– $\lceil \mathsf{B}P_4 \rceil^{\varnothing}_{\mathscr{M}_3} = \{w : \langle w, \langle w, \varnothing \rangle \rangle \in \|\mathsf{B}P_4\|_{\mathscr{M}_3}\}$

$= \{w : \langle w, \langle w, \varnothing \rangle \rangle \in W \times (W \times 2^{\{w_0, w_1\}})\} = W;$

– $\lceil \mathsf{B}P_4 \rceil^{\{w_0\}}_{\mathscr{M}_3} = \{w : \langle w, \langle w, \{w_0\} \rangle \rangle \in \|\mathsf{B}P_4\|_{\mathscr{M}_3}\}$

$= \{w : \langle w, \langle w, \{w_0\} \rangle \rangle \in W \times (W \times 2^{\{w_0, w_1\}})\} = W.$

- $\lceil \neg \mathsf{B}P_4 \rceil^{\{w_0\}}_{\mathscr{M}_3} = \{w : \langle w, \langle w, \{w_0\} \rangle \rangle \in \|\neg \mathsf{B}P_4\|_{\mathscr{M}_3}\}$

$= \{w : \langle w, \langle w, \{w_0\} \rangle \rangle \in (W \times \mathscr{C}) \setminus (W \times (W \times 2^{\{w_0, w_1\}}))\} = \varnothing.$

Exercise

Show:

- $\lceil \neg \varphi \rceil^{i_c}_{\mathscr{M}} = W \setminus \lceil \varphi \rceil^{i_c}_{\mathscr{M}}$;
- $\lceil \varphi \wedge \psi \rceil^{i_c}_{\mathscr{M}} = \lceil \varphi \rceil^{i_c}_{\mathscr{M}} \cap \lceil \psi \rceil^{i_c}_{\mathscr{M}}$;
- $\lceil \mathsf{A}\varphi \rceil^{i_c}_{\mathscr{M}} = \lceil \varphi \rceil^{i_c}_{\mathscr{M}}$;
- $\lceil \mathsf{B}\varphi \rceil^{i_c}_{\mathscr{M}} =$
  - $W \equiv i_c \subseteq \lceil \varphi \rceil^{i_c}_{\mathscr{M}}$;
  - $\varnothing \equiv i_c \nsubseteq \lceil \varphi \rceil^{i_c}_{\mathscr{M}}$.

These results are quite nice because they allow us to ignore the conceptual acrobatics required by thinking about multidimensional semantic values.

Presupposition

Note that we can now characterize presupposition in the following way:

- φ presupposes ψ just if for all $i_c, \mathscr{M}$, $\lceil \varphi \rceil^{i_c}_{\mathscr{M}} \subseteq \lceil \psi \rceil^{i_c}_{\mathscr{M}}$;
- The sentences in Δ collectively presuppose ψ just if for all $i_c, \mathscr{M}$, $\lceil \bigwedge \Delta \rceil^{i_c}_{\mathscr{M}} \subseteq \lceil \psi \rceil^{i_c}_{\mathscr{M}}$;
  Where $\bigwedge \Delta$ is the sentence that conjoins every sentence in Δ.

10.6 Assertion

The contrast among the clauses presenting the propositional meanings of the sentences of L_3 between the clause for $\mathsf{B}\varphi$ and the rest is frequently put by saying that belief sentences do not 'convey information' but instead 'test the context'.

To see the significance of this observation, it will be helpful to have in mind a sort of 'formal pragmatics'. Suppose that the aim of conversation, or inquiry more generally, is to learn stuff. We can model this by thinking of a process of inquiry as a sequence of contexts $\{c_0, \ldots, c_n\}$ such that $i_{c_0} \supseteq \ldots \supseteq i_{c_n}$. Our initial picture of the world has a certain vagueness to it; we sharpen it successively and monotonically at each stage.

In context c_j, the body of information i_{c_j} relevant to that context is referred to as the 'common ground' of that context: in effect, that which everyone who has a stake in the inquiry is agreeing to take for granted. In a solipsistic situation, we may think of the 'common ground' as the subject's beliefs; in a social situation, the common ground may be an information state corresponding to the disjunction of the belief states of the various participants. These are perhaps the most straightforward understandings of the notion of common ground, but the apparatus is useful in other cases as well: in particular, for the important case of reasoning within a hypothesis, or for accepting a claim arguendo, or for the development of a plausible fiction, or ... .

That's the synchronic story. The diachronic story concerns the source and nature of inter-stage transitions. Roughly, a transition from stage to stage is the result of 'conditionalization' on the inter-stage assertion: coming to ignore every possibility incompatible with the content of that assertion. A bit more sharply (suppressing explicit reference to a model):

- **The essential effect:** suppose that for $j \in \{1, \ldots, n\}$, φ_j is asserted in c_{j-1}, and in each case the assertion is accepted.

  In that case, $i_{c_{j+1}} = i_{c_j} \cap \lceil \varphi_j \rceil^{i_{c_j}}$.

This is our first rule of 'formal pragmatics'.

In a B-free language, the reference to a background context against which propositions are assigned to sentences can be suppressed, the essential effect can be stated more simply:

- $i_{c_{j+1}} = i_{c_j} \cap \lceil \varphi_j \rceil$; accordingly, $i_{c_n} = i_{c_0} \cap \bigcap_{j=1}^{n} \lceil \varphi_j \rceil$.

Things look a bit different when we account for B, in a way that will require us to add an important modification to our formal pragmatics. Suppose that at a certain stage of inquiry c_j, someone says 'well, certainly φ'. Here there are two possibilities:

- If $i_{c_j} \subseteq \lceil \varphi \rceil^{i_{c_j}}$, then $i_{c_{j+1}} = i_{c_j} \cap \lceil \mathsf{B}\varphi \rceil^{i_{c_j}} = i_{c_j} \cap W = i_{c_j}$: the common ground remains unchanged by the assertion.
- If $i_c \nsubseteq \lceil \varphi \rceil^{i_c}$, then $i_{c_{j+1}} = i_{c_j} \cap \lceil \mathsf{B}\varphi \rceil^{i_{c_j}} = i_{c_j} \cap \varnothing = \varnothing$.

In the former case, we were already certain that φ, so we nod our heads and go on: the common ground remains unchanged by the assertion. In the latter case, we were not certain that φ. What then? If the assertion is accepted, then in accord with the essential effect, the common ground will contract to the null set.

Well what would it be for that to happen? If the common ground consists of the worlds we are willing to accept as maybe actual, then if the common ground is empty, there are *no* worlds we think are actual. There are a lot of problems with this situation. It is of course manifest that at least one world is actual: this one! And inquiry would seem to come to an end: if the aim of inquiry is to sharpen our picture of the world, well you can't get any sharper than the null set. And finally if what is believed are the propositions containing the common ground, then since every set contains the null set, in this situation every proposition (not to mention its negation) is believed.

This situation is an example of what we call a 'defective context'. Let us amend our formal pragmatics to incorporate the following absolute prohibition:

- **Defectiveness avoidance:** For all c, c is nondefective; in particular, $i_c \neq \varnothing$.

Note an 'escape clause' in the characterization of the essential effect: in order for an assertion to *bring about* the essential effect, it needs to be *accepted*. (This can be seen in the more anodyne case in which someone tries to advance some information that is understood as already ruled out: 'and Saddam's nuclear program was progressing apace, so—'; 'hey wait a second, that's false'. In that case the assertion was just not accepted.) So an assertion of $\mathsf{B}\varphi$ in a context in which φ is not accepted need not 'crash' the context: the assertion may instead be rejected.

Predictively, this seems inadequate: in most reasonable conversations, φ and 'certainly φ' are interchangeable. So we need some story for how assertion of the latter can not just fail to crash a context in which φ is uncertain, but even result in conditionalization of the context on φ.

The resource here is that people (people who are not philosophers or hard-core geeks of other tribal affiliations, that is) tend to be highly lenient in shifting the common ground around so as to avoid rejecting assertions. This is the phenomenon of *accommodation*. What exactly accommodation involves is a matter of unresolved debate, but here is a classic statement from Lewis:

> If at time t something is said that requires presupposition P to be acceptable, and if P is not presupposed just before t, then—ceteris

paribus and within certain limits—presupposition P comes into existence at t.

In our terms, this can be put as the following amendment to the essential effect:

- **Accommodation:** suppose the following:
  - φ presupposes ψ;
  - $i_{c_j} \cap \lceil \varphi \rceil^{i_{c_j}}$ is defective;
  - $i_{c_j} \cap \lceil \psi \rceil^{i_{c_j}} \cap \lceil \varphi \rceil^{i_{c_j} \cap \lceil \psi \rceil^{i_{c_j}}}$ is nondefective.

  Then, if φ is asserted and accepted in c_j,

$$i_{c_{j+1}} = i_{c_j} \cap \lceil \psi \rceil^{i_{c_j}} \cap \lceil \varphi \rceil^{i_{c_j} \cap \lceil \psi \rceil^{i_{c_j}}}.$$

That formula is, of course, barely discriminable from complete gibberish, but it expresses a doctrine with a lot of intuitive pull. What it says is that if the essential effect of accepting an assertion would crash the context, then we: (1) search around for a presupposition of that assertion that (a) would not crash the context and (b) an intermediate context that updates the original context with the presupposition (via the essential effect) would not *be* crashed by the assertion; (2) make the intermediary update of the context with the presupposition via the essential effect; and (3) update the intermediary context with the original assertion (understood against the intermediary context) via the essential effect.

If we want a simpler way to express the accommodation rule, we could do it this way:

- Let $j \dot{+} \psi = i_{c_j} \cap \lceil \psi \rceil^{i_{c_j}}$.
- **Accommodation:** suppose the following:
  - φ presupposes ψ;
  - $i_{c_j} \cap \lceil \varphi \rceil^{i_{c_j}}$ is defective;
  - $j \dot{+} \psi \cap \lceil \varphi \rceil^{j \dot{+} \psi}$ is nondefective.

  Then, if φ is asserted and accepted in c_j,

$$i_{c_{j+1}} = j \dot{+} \psi \cap \lceil \varphi \rceil^{j \dot{+} \psi}.$$

Now suppose that we follow the accommodation rule in order to update the P_4-uncertain context c_j with $\mathsf{B}P_4$: suppose $i_{c_j} = W$. The update rolls out as follows:

- $\mathsf{B}P_4$, as we have seen, presupposes P_4;
- $i_{c_j} \cap \lceil \mathsf{B}P_4 \rceil^W$ is, as we have seen, defective;
- But $j + P_4 \cap \lceil \mathsf{B}P_4 \rceil^{j+P_4}$ is nondefective: the intermediate update returns the set of all P_4-worlds (namely $2^{\{w_0,w_1\}}$); against the intermediate update, $\mathsf{B}P_4$ expresses the proposition W; and intersecting a nondefective context with W results in a nondefective context (namely the original context);
- So the information state of the follow-on context, $i_{c_{j+1+}}$, is $2^{\{w_0,w_1\}}$.

This is of course possible because $\mathsf{B}\varphi$ expresses a *different* proposition depending on which body of information it is evaluated against.

It is worth reiterating that this particular treatment of accommodation is highly preliminary and based on the (more semantical) Lewis approach rather than the (more pragmatical) Stalnaker approach: alternatives will of course generate differing empirical predictions. Obviously introducing variant understandings of presupposition would have the same effect.

11 Update semantics

11.1 Basics

Perhaps rather than understanding meanings as primarily representational entities and deriving pragmatic effects via the sorts of principles under consideration, we could flip the perspective: take linguistic meaning as primarily pragmatic and derive representational properties later if at all. Flipping the perspective in this way moves us from our previous 'static' approach to the in many ways rather different 'dynamic' or 'update' approach.

Here is what I mean by saying that meaning is primarily pragmatic: the object with which we associate the meaning of φ is not a proposition (or even a proposition relative to a body of information) but rather an 'update rule'. This would be a *function* characterizing the effect on a context of accepting an assertion of φ: a function mapping a prior or background context into a posterior or resultant or updated context; a *context change potential*. We could then represent φ's meaning, $[\varphi]$, as a rule, or a set of ordered pairs, or whatever.

This approach would not necessarily sever the connection between linguistic meaning and representation. For all the approach at its most bare-bones cares, perhaps contexts can or must be characterized partly or solely in terms of representation. And indeed, we will begin by investigating approaches on which contexts are propositions understood, as before, as sets of worlds.

Some notational peculiarity should not trip us up. Ordinarily functions are written as prefixes of their arguments: $f(x)$ is the function f applied to the argument x. But since conversation is a process in time, reading is a process in time, and we read from left to right, update theorists have a fondness for the opposite approach: $c[\varphi][\psi][\rho]$ is the context resulting from successively applying to c the meanings of φ, ψ, and ρ.

In general, characterizing the logical form of the context-change potential function is extremely straightforward:

- $[\cdot] : L \mapsto \mathscr{C} \times \mathscr{C}$;
  or perhaps: $(\cdot)[\cdot] : L \times \mathscr{C} \mapsto \mathscr{C}$.
- An explicitly model-relative formulation would run:
  $[\cdot]_{\mathscr{M}} : L \mapsto \mathscr{C} \times \mathscr{C}$.

Changing the language to introduce or eliminate complexity does not require reconceiving our very conception of what meaning is: all the complexity can be bundled into our characterization of context. Perhaps this carries some theoretical advantage.

11.2 Update rules for L_0 + B

Compositional update rules for $L_0 + \mathsf{B}$ look like this (suppressing explicit relativization to a model):

- For $\varphi \in \Pi_0, i_c[\varphi] = i_c \cap V(\varphi)$;
- $i_c[\neg\psi] = i_c \setminus i_c[\psi]$;
- $i_c[\psi \wedge \rho] = \ldots$
  - $i_c[\psi][\rho]$ ('update' conjunction); or perhaps
  - $i_c[\psi] \cap i_c[\rho]$ ('classical' conjunction)
- $i_c[\mathsf{B}\psi] = \ldots$
  - i_c just if $i_c[\psi] = i_c$;
  - $\varnothing$ just if $i_c[\psi] \neq i_c$.

Comments:

- Observe the familiar use of recursion; here used to characterize update potentials, where the base clauses are, again, those for sentences in Π_0.

- The predictions of the rules for $\neg$ and B are the same as those from our earlier languages, supplemented with the rule of the essential effect.

- The rule for 'classical' conjunction is also the same.

  However the rule for 'update' conjunction makes different predictions: it is not the case in general that $i_c[\neg \mathsf{B} P_4][P_4] = i_c[P_4][\neg \mathsf{B} P_4]$: if i_c is uncertain about P_4, then the former value is $i_c \cap V(P_4)$, while the latter is $\varnothing$.

  Update conjunction is really a very unfamiliar beast. Expressing the analogous idea in the static framework would require a clause like this one:

  – $\ulcorner \varphi \wedge \psi \urcorner^{i_c}_{\mathscr{M}} = \ulcorner \varphi \urcorner^{i_c}_{\mathscr{M}} \cap \ulcorner \psi \urcorner^{i_c \cap \ulcorner \varphi \urcorner^{i_c}_{\mathscr{M}}}_{\mathscr{M}}$.

  One thing to notice about this clause is that it *effectively takes a stand on a pragmatic matter*: it requires the interpretation of the second conjunct to depend not just on the context, but on *how the context should change under the incorporation of the first conjunct*—via the essential effect, rather than by some other means. This breaks the barrier between semantics and pragmatics, in effect converting the approach into an update system.

  Next, it is unlikely that the above formula explains much about our psychological process for grasping conjunction: add a third conjunct, and we immediately reach a risibly complex formula concerning values under hypothetical contexts embedded three deep.

  A final point is that the information-relative proposition function $\ulcorner \cdot \urcorner^{\cdot}$ is defined solely in terms of the semantic value function $\| \cdot \|$, where we regarded it as an implicit desideratum on the proposition function that, for sentences of L_0, the point in the derivation in which the proposition function is applied does not matter (thus the exercise in section 7.5). As should be obvious, this requires that update conjunction can't be the expression described by $\| \cdot \|_{\mathscr{M}_0}$: indeed, the resources of $\mathscr{M}_0$ are too impoverished to define update conjunction. But without classical conjunction, we have no apparatus for building sentences containing more than one atomic letter (natural language 'if' and 'or' are clearly too complex for $\| \cdot \|_{\mathscr{M}_0}$).

  If update conjunction is 'real' conjunction, the static approach is on very shaky ground as a correct story about 'real' logical complexity.

11.3 Discourse entailment

We can define a range of notions of entailment in an update framework. The most straightforward is the following:

- φ discourse-entails ψ just if for all c, $\mathscr{M}$, $c[\varphi]_{\mathscr{M}}[\psi]_{\mathscr{M}} = c[\varphi]_{\mathscr{M}}$.

A bit more expansively, φ discourse-entails ψ just if no matter what information we start with, in updating it with φ, we have in effect thereby already updated it with ψ.

Patterns of discourse-entailment with epistemic modals and negation shake out as follows:

- $\varphi \vdash \mathsf{B}\varphi$ (obviously);
- $\mathsf{B}\varphi \vdash \varphi$ (perhaps less obviously: but in the event that the first update crashes the context, the second update won't uncrash it);
- $\neg\varphi \vdash \mathsf{B}\neg\varphi$;
- $\neg\varphi \vdash \neg\mathsf{B}\varphi$;
- $\mathsf{B}\neg\varphi \vdash \neg\varphi$;
- $\neg\mathsf{B}\varphi \nvdash \neg\varphi$—nor, of course, does it entail φ.

The story for multipremiss discourse-entailment looks different. In update semantics, order matters. Accordingly, the notion of entailment by a *set* makes no sense. The relevant notion is rather entailment by a *sequence*, which works as follows:

- The sentences in the sequence $\langle \psi_1, \ldots, psi_n \rangle$ discourse-entail φ just if for all c, $c[\psi_1]\ldots[\psi_n][\varphi] = c[\psi_1]\ldots[\psi_n]$.

Note that our notion of presupposition yields different predictions than our notion of discourse-entailment:

- Consider the following arguments:
  - $\neg\mathsf{B}\varphi$; φ: so ψ;
  - φ; $\neg\mathsf{B}\varphi$: so ψ.
- By the standard of presupposition, each is valid: $\bigwedge\{\neg\mathsf{B}\varphi, \varphi\} = \bigwedge\{\varphi, \neg\mathsf{B}\varphi\} = \neg\mathsf{B}\varphi \wedge \varphi$; and in every c, $\lceil \neg\mathsf{B}\varphi \wedge \varphi \rceil^c = \varnothing$; so, since whatever ψ may be, $\varnothing \subseteq \lceil \psi \rceil^c$, the premisses collectively presuppose the conclusion.

- By contrast, by the standard of discourse-entailment, only the latter is generally valid. Let $i_c = W$. Then $c[\neg \mathsf{B}\varphi] = c$, so $c[\neg \mathsf{B}\varphi][\varphi] = c[\varphi] \neq c[\varphi][\psi]$, for some φ and ψ.

The remainder of this handout presents a range of applications of the update approach.

12 Conditionals

A relatively widely explored arena of application for the update approach is to the semantics of conditionals.

12.1 Properties of conditionals

We observe the following phenomena:

- **Indicative-subjunctive contrast:**
  - If Oswald didn't shoot Kennedy, someone else did (true);
  - If Oswald hadn't shot Kennedy, someone else would have (false, at least if, like me, you do not regard Warrenism as obviously false).
- **Latitude in the consequent:**
  - If you see Keyser Sose, you will die soon;
  - If you see Keyser Sose, it is necessary that you will die soon;
  - If you see Keyser Sose, you might have time to warn the others;
  - If you see Keyser Sose, what is he wearing?
  - If you see Keyser Sose, run!
- **Valitude in the antecedent:**
  - If you will die soon, you see Keyser Sose;
  - ? If it is necessary that you will die soon, you see Keyser Sose;
  - ?? If you might have time to warn the others, you see Keyser Sose;
  - * If what is he wearing?, you see Keyser Sose;
  - * If run!, you see Keyser Sose.
- **Evidential affiliation** (Kolodny and MacFarland):

- If the miners are in shaft A, we should flood shaft B;
- If the miners are in shaft B, we should flood shaft A;
- The miners are either in shaft A or shaft B;
- $\not\vdash$ Either we should flood shaft B or we should flood shaft A.

- **Long-distance cut:**
  - If you see Keyser Sose, you will die soon; ...; Look, Keyser Sose! $\vdash$ You will die soon;
  - If you see Keyser Sose, it is necessary that you will die soon; ...; Look, Keyser Sose! $\vdash$ It is necessary that you will die soon;
  - If you see Keyser Sose, you might have time to warn the others; ...; Look, Keyser Sose! $\vdash$ You might have time to warn the others;
  - If you see Keyser Sose, what is he wearing?; ...; Look, Keyser Sose! $\vdash$ What is he wearing?;
  - If you see Keyser Sose, run!; ...; Look, Keyser Sose! $\vdash$ Run!

- **Ramsey test:**
  - If you see Keyser Sose, you will die soon. Really? Let's suppose I see Keyser Sose ... in which case, I will die soon. So OK.
  - If you had seen Keyser Sose, you would have died quickly. Really? Let's suppose I had seen Keyser Sose ... in which case, I would have died soon. So OK.
  - If you see Keyser Sose, it is necessary that you will die soon. Really? Let's suppose I see Keyser Sose ... in which case, it is necessary that I will die soon. So OK.
  - If you see Keyser Sose, you might have time to warn the others. Really? Let's suppose I see Keyser Sose ... in which case, I might have time to warn the others. So OK.
  - If you see Keyser Sose, what is he wearing? Really? Let's suppose I see Keyser Sose ... but his outfit is the last thing on our mind! So fuhgeddaboutit!
  - If you see Keyser Sose, run! Really? Let's suppose I see Keyser Sose ... in which case, run! So OK.

12.2 Indicatives and subjunctives

Here's a nice toy update-style theory for the indicative and subjunctive conditional (suppressing some of the complexity in Will Starr's story):

- Indicative: $i_c[\mathsf{if}\ \varphi, \psi] = \ldots$
  - i_c just if $i_c[\varphi][\psi] = i_c[\varphi]$;
  - $\varnothing$ just if $i_c[\varphi][\psi] \neq i_c[\varphi]$.
- Subjunctive: $i_c[\mathsf{if\ were}\ \varphi, \psi] = \ldots$
  - i_c just if $i_c[\mathsf{were}\ \varphi][\psi] = i_c[\mathsf{were}\ \varphi]$;
  - $\varnothing$ just if $i_c[\mathsf{were}\ \varphi][\psi] \neq i_c[\mathsf{were}\ \varphi]$;

  Where
  - $i_c[\mathsf{were}\ \varphi]$ = the set resulting from replacing every $\neg\varphi$ world in i_c with the φ world most similar to it.

This proposal can be cast into the 'test' form of the static semantic theory developed for B in 7.5, as follows:

- $\lceil \mathsf{if}\ \varphi, \psi \rceil^{i_c} =$
  - $W \equiv \lceil \varphi \rceil^{i_c} \subseteq \lceil \psi \rceil^{i_c + \varphi}$;
  - $\varnothing \equiv \lceil \varphi \rceil^{i_c} \not\subseteq \lceil \psi \rceil^{i_c + \varphi}_{\mathscr{M}}$.

For bookkeeping purposes, let $L_4 = L_3 + \mathsf{if} + \mathsf{were}$.

12.3 Advantages of the proposal

This theory allows us to sketch explanations for several target phenomena:

- **Oswald-Kennedy:** Let's suppose that every world in the context set is an *Oswald shot Kennedy* world, and therefore a *someone shot Kennedy* world.
  - **The indicative:**
    Updating on $\neg$*Oswald shot Kennedy* creates a defective context. Perhaps we avoid this by altering the context set to include some $\neg$*Oswald shot Kennedy* worlds, but in such a way as to preserve as much of our knowledge as is compatible with this: for example, every world remains a *someone shot Kennedy* world. (This is very hand-wavy.)

Updating the resultant on *someone distinct from Oswald shot Kennedy* leaves it unchanged.

So, the theory predicts, the indicative conditional is acceptable.

– **The subjunctive:**

Replacing every world in the context set with the maximally similar ¬*Oswald shot Kennedy* world, if we are at least somewhat sympathetic to Warrenism, generates a context set including at least some worlds in which Oswald narrowly missed and there was no backup shooter. This context set, therefore, contains some ¬*someone distinct from Oswald shot Kennedy* worlds.

Updating this context set with *someone distinct from Oswald shot Kennedy* yields a distinct context set.

So, the theory predicts, the subjunctive conditional is not acceptable.

- **Long-distance cut for the indicative:**

Suppose *If you see Keyser Sose, you will die soon* is acceptable in c_j. So every *you see Keyser Sose* world in i_{c_j} is a *you will die soon* world.

Further updates result in a context c_k. Now, if $i_{c_k} \subseteq i_{c_j}$, it remains the case that if there are any *you see Keyser Sose* worlds in i_{c_k}, they are *you will die soon* worlds.

So, if we update c_k with *you see Keyser Sose*, the resulting context $i_{c_{k+1}}$ supports *you will die soon*: every world in $i_{c_{k+1}}$ is a *you will die soon* world.

From time to time of course updates are 'destructive': result in possibilities of which we were previously unaware entering the context set. In rare cases in which we abandon earlier certainty, previously acceptable indicative conditionals will not in general be acceptable afterward. In such cases we will observe failures of long-distance cut.

Explaining long-distance cut for other conditionals requires some treatment of other speech acts: later.

- **Ramsey test for indicatives and subjunctives:**

'A conditional is acceptable just if, upon hypothetically adding the antecedent to one's stock of beliefs, the consequent is acceptable'.

Let us suppose that hypothetical and genuine acceptance of an assertion do not differ in their effects on the resulting context set. By this I mean the following. Suppose that c is updated to c^+ and c' is updated to c'^+; that $i_c = i_{c'}$; and that c and c' are both updated solely by 'adding' φ: then $i_{c^+} = i_{c'^+}$:

whether the addition of φ was sincere or hypothetical makes no difference to what is (sincerely or hypothetically) accepted as a result of the addition.

If so, we can ignore the 'hypothetical' aspect of the Ramsey test and cast the discussion solely at the level of context change potential.

- **Indicatives:**
  Left-to-right:
  Suppose that if φ, ψ is acceptable in c. Then $i_c[\varphi][\psi] = i_c[\varphi]$.
  Then every world in $i_c[\varphi]$ is a ψ-world.
  Right-to-left:
  Suppose that every world in $i_c[\varphi]$ is a ψ-world. Then $i_c[\varphi][\psi] = i_c[\varphi]$.
  The theory says that this is what it is for the indicative to be acceptable in c: to fail to crash the context.
- **Subjunctives:**
  Left-to-right:
  Suppose that if $\mathsf{were}\,\varphi, \psi$ is acceptable in c. Then $i_c[\mathsf{were}\,\varphi][\psi] = i_c[\mathsf{were}\,\varphi]$.
  Then every world in $i_c[\mathsf{were}\,\varphi]$ is a ψ-world.
  Right-to-left:
  Suppose that every world in $i_c[\mathsf{were}\,\varphi]$ is a ψ-world. Then $i_c[\mathsf{were}\,\varphi][\psi] = i_c[\mathsf{were}\,\varphi]$.
  The theory says that this is what it is for the subjunctive to be acceptable in c: to fail to crash the context.

- **Entailment and conditionals**

  We have:

  - $\varphi \vdash \psi$ just if for all c, $i_c[\varphi][\psi] = i_c[\varphi]$
  - $i_c[\mathsf{if}\,\varphi, \psi] = \ldots$
    - i_c iff $i_c[\varphi][\psi] = i_c[\varphi]$;
    - $\varnothing$ otherwise.

  Assuming that $\mathsf{if}\,\varphi, \psi$ is acceptable in c just if $i_c[\mathsf{if}\,\varphi, \psi] \neq \varnothing$, we have the following result:

  - $\varphi \vdash \psi$ just if: for all c, if φ, ψ is acceptable in c.

12.4 Ramsey + Moore = God?

Chalmers and Hajek noticed the following interesting paradox (in the course of addressing it we will also note further examples of, and explain, the phenomenon of 'evidential affiliation'):

> According to Moore, if I accept φ but not 'I believe that φ' I am incoherent; and if I accept 'I believe that φ' but not φ I am incoherent.
>
> According to Ramsey, we should accept 'if φ, ψ' just if upon the hypothetical addition of φ to one's stock of beliefs, we hypothetically accept ψ.
>
> So suppose I hypothetically add φ to my stock of beliefs: in the new hypothetical context, I also endorse 'I believe that φ' on pain of incoherence. So right now, I should accept 'if φ, I believe that φ'.
>
> And suppose I hypothetically add 'I believe that φ' to my stock of beliefs: in the new hypothetical context, I also endorse φ on pain of incoherence. So right now, I should accept 'if I believe that φ, φ'.
>
> The former conditional is an 'omniscience' principle: I believe every truth. The latter conditional is an 'infallibility' principle: I believe no falsehoods.
>
> So accepting the Moore and the Ramsey principles commits us to thinking of ourselves as having Godlike epistemic powers.

Some observations:

- Putting things our way: both directions of the 'Moore schema' are valid in L_4. And the Ramsey inference—from $\varphi \vdash \psi$ to $\vdash$ if φ, ψ—is also valid in L_4. So both of these arguments are valid in L_4:
  - $\varphi \vdash \mathsf{B}\varphi$; so if $\varphi, \mathsf{B}\varphi$ is acceptable in every context; so it is a theorem;
  - $\mathsf{B}\varphi \vdash \varphi$; so if $\mathsf{B}\varphi, \varphi$ is acceptable in every context; so it is a theorem.

  If this amounts to affirmation in every context of both omniscience and infallibility, then my possession of Godlike epistemic powers is a theorem of L_4.

  So the question here is whether truth of the two conditionals amounts to omniscience and infallibility in an intuitive sense: as we will see later, when the conditionals are interpreted so that they are theorems, the answer is no.

- Notably, while the contrapositive of infallibility (if $\neg\varphi, \neg \mathsf{B}\varphi$) is a theorem, the contrapositive of omniscience (if $\neg \mathsf{B}\varphi, \neg\varphi$) is not.

  We will pick up on the ramifications of this point later.

- It is easy to prove the existence of a counterexample to the omniscience schema. Either the current population of Russia exceeds 180 million or it does not; and yet I have no opinion one way or the other. Let P be the true one of *the current population of Russia exceeds 180 million* and its negation. Then while P, I do not believe that P.

  There is a notable similarity here to the miners puzzle. Recall:

  - If the miners are in shaft A, we should flood shaft B;
  - If the miners are in shaft B, we should flood shaft A;
  - The miners are either in shaft A or shaft B;
  - $\neg$(Either we should flood shaft B or we should flood shaft A).

  In the present case, the existence proof amounts to the fact that all the following seem acceptable:

  - if $\varphi, \mathsf{B}\varphi$;
  - if $\neg\varphi, \mathsf{B}\neg\varphi$;
  - $\varphi \vee \neg\varphi$;
  - $\neg(\mathsf{B}\varphi \vee \mathsf{B}\neg\varphi)$.

  The difficulty here is that the negation of the fourth claim follows from the first three by reasoning by dilemma: same for the miners.

- Doing the same for infallibility would not work. consider the following alleged aporia:

  - if $\mathsf{B}\varphi, \varphi$;
  - if $\neg \mathsf{B}\varphi, \neg\varphi$;
  - $\mathsf{B}\varphi \vee \neg \mathsf{B}\varphi$;
  - $\neg(\varphi \vee \neg\varphi)$.

  We observe two respects of disanalogy. First, the second line is not an instance of the infallibility schema, and indeed is not a theorem of L_4. Second, the fourth line is a contradiction.

  The result of attempting to fix the first problem is instructive:

- if $\mathsf{B}\varphi, \varphi$;
- if $\mathsf{B}\neg\varphi, \neg\varphi$;
- $\mathsf{B}\varphi \vee \mathsf{B}\neg\varphi$;
- $\neg(\varphi \vee \neg\varphi)$.

The second line is now of the right form to be the infallibility schema, but the third line—the *opinionation schema*—is not true and not a theorem of L_4; indeed, it is just what is denied in proving the existence of a counterexample to the omniscience schema. So the best this strategy can do for us is to show that our infallibility schema is not generally true of an opinionated being in an inconsistent world.

- Reasoning informally, we would like to say something like the following about our worry for omniscience:

  > It is an epistemic possibility that the current population of Russia exceeds 180 million; it is an epistemic possibility that the current population of Russia does not exceed 180 million.
  >
  > And yet I am certain that I have no opinion one way or the other.
  >
  > So shouldn't the following scenario be an epistemic possibility for me? —The current population of Russia exceeds 180 million and I do not believe that it does. (The same, for that matter, goes for the following scenario: the current population of Russia does not exceed 180 million and I do not believe that it does not.)

  Expressing this line of reasoning in L_4 would involve something like this:

  - $\neg\mathsf{B}R \wedge \neg\mathsf{B}\neg R$;
  - $\mathsf{B}(\neg\mathsf{B}R \wedge \neg\mathsf{B}\neg R)$;
  - So: $\neg\mathsf{B}\neg(R \wedge \neg\mathsf{B}R)$.

  To assess what is going on here, let's swap out the Bs for $\Box$es:

  - $\neg\Box R \wedge \neg\Box\neg R$;
  - $\Box(\neg\Box R \wedge \neg\Box\neg R)$;
  - So: $\neg\Box\neg(R \wedge \neg\Box R)$.

  This argument is valid in the weak modal system **K**:

> The first of these is true at w just if the set of worlds accessible from w contains a mix of R- and $\neg R$-worlds—just if w is 'R-contingent'.
>
> The second is true at w just if every world accessible from w is also R-contingent.
>
> And the third is true at w just if at some world v accessible from w, both of these are true: R; at some world accessible from v, $\neg R$.
>
> The third is therefore a valid consequence of the first two in **K**. After all, the model theoretic expression of the third follows from that of the first two by quantificational logic alone, independent of the character of the accessibility relation. For pick an arbitrary world v^* accessible from w at which R: such a world is guaranteed to exist by the R-contingency of w. Then the second premiss tells us that v^* is also R-contingent: so there is some world accessible from it at which $\neg R$. So v^* witnesses the outer existential quantifier in the model theoretic expression of the third claim.

But an analogous argument cannot be given for the validity of the argument expressing our worry for omniscience. The model theoretic expression of *that* argument would run like this:

> The first premiss is true relative to a body of information just if the worlds compatible with that information include both R- and $\neg R$-worlds.
>
> The second is true relative to a body of information just if every world compatible with that information is this way: the worlds compatible with that information include both R- and $\neg R$-worlds. Putting the point less circumspectly: just if the worlds compatible with that information include both R- and $\neg R$-worlds.
>
> The third is true relative to a body of information just if the worlds compatible with that information include worlds of the following sort: worlds which are such that (i) every world in the body of information is an R-world; and (ii) some world in the body of information is a $\neg R$-world.

And of course while plenty of bodies of information satisfy the condition imposed by the first premiss—which is of course the same as the condition imposed by the second premiss—no body of information satisfies the condition imposed by the third premiss.

Putting things informally, the disanalogy could be thought of in this way. Let φ be an elemental sentence and assume a fixed model. Then whether φ is true at a world w is entirely *internal to the character of modal space*. It is determined by how things are *at* w. Whether $\Box\varphi$ is in turn true is still internal to the character of modal space. The dependence is less 'local' to w, in the sense that the truth-value depends in turn on how things are at the worlds w sees, but that's it. As a result, it makes sense to ask, *of* w, whether φ is true at it, and whether $\Box\varphi$ is true at it.

By contrast, whether $\mathsf{B}\varphi$ is true 'at w' is *not* internal to the character of modal space. The status of this sentence depends on the scope of our attention: if the class of worlds we are not ignoring includes only φ-worlds, the sentence is true simpliciter (and therefore true, in a somewhat degenerate and unhelpful sense, 'at w'); otherwise it is false simpliciter (and therefore false, in a somewhat degenerate and unhelpful sense, 'at w').

Accordingly, the style of reasoning behind our worry for omniscience, though alluring, is mistaken:

> As I survey the region of modal space I am not ignoring, I notice worlds at which R and worlds at which $\not R$. I myself am uncertain whether R. So surely this uncertainty (as it were) filters downward, into the epistemic possibilities themselves. So the worlds at which R are also worlds in which I am uncertain whether R! And if so, there are worlds in which (R and I am uncertain whether R). That kind of world is a counterexample to omniscience!

But to focus in on the worlds at which R is to destroy my uncertainty whether R. When I start ignoring $\neg R$, I am no longer failing to ignore any $\neg R$-worlds. And when I do so, $\mathsf{B}R$ becomes true of me.

Obviously there is no analogy to the behavior of $\Box$ here. Ignoring the $\neg R$-worlds has no influence over which worlds $\Box R$ is true at. The facts determining the semantic properties of $\Box R$ are 'immanent' in modal space; but the facts determining the semantic properties of $\mathsf{B}R$ are not. It therefore makes sense to think of $\Box R$ as filtering down from a full set of worlds to its members; but it makes no sense to think the same of $\mathsf{B}R$.

We can put this by saying that $\Box\varphi$ is 'dissective' in Goodman's sense; $\mathsf{B}\varphi$ is not. (For the time being, let's treat this as a suggestive analogy pending more rigorous development.) By the claim that φ is dissective, I mean this: if, focusing our attention on a certain region of modal space, φ is true at

every world in that region: then, for any subregion of that region, focusing our attention on that subregion, φ is true at every world in that subregion.

Conversely, we can say that $\Box\varphi$ is 'persistent' in Goodman's sense; $\mathsf{B}\varphi$ is not. By the claim that φ is persistent, I mean this: if, focusing our attention on a certain region of modal space, φ is true at every world in that region: then, for any superregion of that region, focusing our attention on that superregion, φ is true at some world in that superregion.

Dissectiveness and persistence are very intuitive features of properties that are intrinsic to regions. (It would be cool if this could be put formally and proved.) They are not, in general, features of properties that are extrinsic of regions, however.

- This failure of the world-relative truth-values of certain sentences to be either dissective or persistent over modal space explains the nonvalidity of reasoning by dilemma, namely the following schema:

  - if ψ, φ
  - if ρ, φ
  - $\psi \vee \rho$
  - $\vdash \varphi$

  We can see the difficulty nondissectiveness and nonpersistence pose for dilemma by reasoning informally as follows.

  Let's grant that, focusing our attention on the region R of modal space, every world in R is either $\psi \vee \rho$. And let's grant that, focusing our attention on R_ψ, the subregion of R containing exactly the worlds at which ψ, every world in R_ψ is α and therefore $\varphi = \alpha \vee \beta$. And let's grant that, focusing our attention on R_ρ, the subregion of R containing exactly the worlds at which ρ, every world in R_ρ is β and therefore $\varphi = \alpha \vee \beta$.

  Granting all this, *nothing follows* concerning whether, focusing our attention on the region R of modal space, every world in R is φ. If α (β) is nonpersistent, then it may be that, though, considering R_ψ (R_ρ), α (β) is true at every world in it—still, considering R, neither α nor β is true at *any* world in it: so that φ is not true at any world in it, either.

- We could resolve the miners puzzle along similar lines by postulating a suitable relationship between 'we should flood shaft B' and 'B the miners are in shaft A'.

  For example, perhaps 'we should flood shaft B' means something like:

– $(\exists\Phi)(\exists a)(\mathsf{B}\Phi \wedge \circledcirc(\Phi, a, \mathfrak{F}_{\mathrm{B}}))$,

Where $\circledcirc(\Phi, \mathfrak{F}_{\mathrm{B}})$ means something like 'in light of the fact that Φ and the aim a, here's the optimal thing to do: $\mathfrak{F}_{\mathrm{B}}!$ (namely, flood shaft B!)'. (We are using German letters to express imperatives.)

That's not exactly right (indeed, more on this is to follow), but it certainly explains the present failure of reasoning by dilemma.

- B is exceedingly sensitive to the context that is currently 'live'. Conditionalizing a context on φ, whether hypothetically or 'for real', results in a context with respect to which $\mathsf{B}\varphi$ is true. If this were not so, if $\mathsf{B}\varphi, \varphi$ and if $\varphi, \mathsf{B}\varphi$ would not be theorems.

  But omniscience is not, intuitively, the following property: taking φ for granted—being such that φ is taken for granted. Omniscience should be more stable, more closely tied to what one 'really' believes than to what is taken for granted, even merely hypothetically. Omniscience is more like this property: taking φ for granted—being such that one *really* believes φ.

  This can be brought out by considering a language L_5 which renders the B operator more flexible: so that it is not bound to the body of information foremost in a context, but can 'float' to other bodies of information. This would allow us to say, even under the hypothesis that φ, that we do not *really* believe that φ. The net effect would be a bit like the introduction of 'double indexing' with respect not just to worlds, but also to bodies of information.

  Here is a provisional sketch of the view to be developed. What we want is to be able to preserve the Moore-schemata while invalidating their more troublesome siblings:

  – if $\varphi, \mathsf{B}_\bigstar\varphi$;
  – if $\mathsf{B}_\bigstar\varphi, \varphi$.

  The idea here is that B, unadorned, continues to perform the sort of 'test' already noted on a certain privileged body of information closely tied to the evaluation of conditionals. However, when subscripted, B shifts its locus of duty over to other bodies of information. Since these will in general have no affiliation with the conditional, there need be no alignment between the consequent and antecedent of the troublesome sort in regard to the complement of B.

Under this analysis, the subscript ★ plays a special role: it is keyed to the body of information that is 'really' taken for granted in the context as opposed to taken for granted merely hypothetically. In that sense, $\mathsf{B}_\bigstar\varphi$ corresponds more closely to an ordinary belief avowal than does $\mathsf{B}\varphi$.

Let us revise our definition of context so that $\mathscr{C} = W \times (2^W \times (2^W \times \mathbb{N} \mapsto 2^W))$: for a particular context c, $c = \langle w_c, i_c, \langle r_c, n^1_c, n^2_c, \ldots\rangle\rangle$. Intuitively, r_c is the body of information really believed in c, as distinguished from i_c, the body of information which is 'live' in c; while the various n^j_c are various other bodies of information kept track of.

Explicit definition of the semantic value function is becoming excessively laborious, so we will scrap it and go straight to explaining meaning in terms of our notion of proposition expressed relative to a context.

Notational change: instead of relativizing to a single body of information, we now relativize to a full context, making explicit how the individual bodies of information in the context interact with proposition expressed:

- $\lceil\varphi\rceil^c = \{w : (\exists\Phi, f)\langle w, \langle w, i_c, \langle\Phi, f\rangle\rangle\rangle \in \|\varphi\|\}$;
- $c \dotplus \varphi = \langle w_c, i_c \cap \lceil\varphi\rceil^c, \langle r_c, n^1_c, n^2_c, \ldots\rangle\rangle$.

Our characterizations of our logical vocabulary carry over without significant changes:

- $\lceil\neg\varphi\rceil^c = W \setminus \lceil\varphi\rceil^c$;
- $\lceil\varphi \wedge \psi\rceil^c = \lceil\varphi\rceil^c \cap \lceil\psi\rceil^c$;
- $\lceil\mathsf{A}\varphi\rceil^c = \lceil\varphi\rceil^c$;
- $\lceil\mathsf{B}\varphi\rceil^c =$
  - $W \equiv i_c \subseteq \lceil\varphi\rceil^c$;
  - $\varnothing \equiv i_c \nsubseteq \lceil\varphi\rceil^c$.
- $\lceil\mathsf{if}\,\varphi, \psi\rceil^c =$
  - $W \equiv i_{c\dotplus\varphi} \subseteq \lceil\psi\rceil^{c\dotplus\varphi}$;
  - $\varnothing$ otherwise.

Here comes a refinement of our initial sketch of our resolution. In order to avoid multiplying senses of 'believes', we will analyze $\mathsf{B}_\bigstar$ into a contribution due to ordinary old B and a sort of 'index-shifting' operator. We will call this operator $\mathsf{FROM}^{(\cdot)}$, where the permissible occupants of the subscript position are any of the singular terms in the sequence $\{\bigstar, s_1, s_2, \ldots\}$. These expressions interact with the rest of the language as follows:

- $\lceil \mathsf{FROM}^{\bigstar} \varphi \rceil^c =$
  - $\lceil \varphi \rceil^{c/\bigstar}$; where
  - $c/\bigstar = \langle w_c, r_c, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$;
- $\lceil \mathsf{FROM}^{s_j} \varphi \rceil^c =$
  - $\lceil \varphi \rceil^{c/s_j}$;
  - $c/s_j = \langle w_c, n_c^j, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$;

The effect of these definitions is to allow us to distinguish what one is taking for granted from what one *really* believes. Our thought is that ordinary uses of 'if φ, I believe that φ' have two distinct logical forms:

- if $\varphi, \mathsf{B}\varphi$;
- if $\varphi, \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi$.

The first of these is a theorem, the second is not; but the second better captures what one would have to accept in order to take oneself to be omniscient.

To see this, consider the conditional if $\varphi, \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi$. Relative to c, this 'tests' for whether $i_{c\dot{+}\varphi} \subseteq \lceil \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi \rceil^{c\dot{+}\varphi}$. Now, $\lceil \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi \rceil^{c\dot{+}\varphi}$ expresses the proposition $\lceil \mathsf{B}\varphi \rceil^{c\dot{+}\varphi/\bigstar}$; and this, in turn, 'tests' for whether $r_c \subseteq \lceil \varphi \rceil^{c\dot{+}\varphi/\bigstar}$. And $\lceil \varphi \rceil^{c\dot{+}\varphi/\bigstar}$, in turn, is determined as follows:

- $c \dot{+} \varphi = \langle w_c, i_c \cap \lceil \varphi \rceil^c, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$; and
- $d/\bigstar = \langle w_d, r_d, \langle r_d, n_d^1, n_d^2, \ldots \rangle \rangle$; so that
- $c \dot{+} \varphi/\bigstar = \langle w_c, r_c, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$; and of course
- $\lceil \varphi \rceil^d = \{w : (\exists \Phi, f) \langle w, \langle w, i_d, \langle \Phi, f \rangle \rangle \rangle \in \|\varphi\|\}$; so that
- $\lceil \varphi \rceil^{c\dot{+}\varphi/\bigstar} = \{w : (\exists \Phi, f) \langle w, \langle w, r_c, \langle \Phi, f \rangle \rangle \rangle \in \|\varphi\|\}$;
- Which is, in effect, the proposition expressed by φ relative to a context c' in which the salient information is not the proposition i_c but rather r_c.

Let's set φ to something contingent, like P_4, in which case this proposition is $\{w_0, w_1\}$. And let's suppose we are *really* uncertain whether P_4 is the case: that $r_c = \{w_1, w_2\}$. In that case, $r_c \not\subseteq \lceil \varphi \rceil^{c\dot{+}\varphi/\bigstar}$. And in that case, $\lceil \mathsf{B}\varphi \rceil^{c\dot{+}\varphi/\bigstar} = \lceil \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi \rceil^{c\dot{+}\varphi} = \varnothing$. On the assumption that $i_c = r_c$, $i_{c\dot{+}\varphi} = \{w_1\}$; and so $i_{c\dot{+}\varphi} \not\subseteq \lceil \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi \rceil^{c\dot{+}\varphi}$, in which case the conditional if $\varphi, \mathsf{FROM}^{\bigstar} \mathsf{B}\varphi$ is not true.

Let's pause to reflect on what just happened: we advanced an interpretation of 'I believe that P' which does *not* support the omniscience direction of the Moore conditional. Indeed, I'm inclined to think that the Moore conditional read as $\mathsf{if}\,\varphi, \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi$ provides a much more plausible understanding of what it would be for me to be omniscient than does its earlier interpretation as $\mathsf{if}\,\varphi, \mathsf{B}\varphi$. This does not, however, undermine the Moore inferences as providing the most basic story about our notion of belief: testing of the live information coordinate of a context is inescapable as a part of the story.

Note that we still preserve the following inference for discourse-initial contexts: $\varphi \vdash \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi$. After all, if I simply take φ for granted in c, where c is not hypothetical but 'unqualified', I conditionalize not just i_c but also r_c on φ; so with respect to c, $\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi$ is acceptable.

Note that we can use our apparatus to characterize what it would be for somebody else—Cheney, for example—to be omniscient. Letting s_{666} name Cheney, the doctrine of Cheney's omniscience would be this:

- $\mathsf{if}\,\varphi, \mathsf{FROM}^{s_{666}}\mathsf{B}\varphi$.

What I affirm there is something like this. Assume whatever you like, the fact remains: Cheney knows all!

- In a certain sense, facts about belief are 'outside of the world': the semantic properties of $\mathsf{B}\varphi$ are not, as we have seen, determined by the intrinsic character of modal space. If this is true for me, it is true for everyone. When I shift to Dick Cheney's perspective, I inhabit another position 'outside the world'. That is, perhaps, a congenial perspective of this system.

- Can our technique be used to resolve the paradoxical infallibility consequence? If so, the conditional $\mathsf{if}\,\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi, \varphi$ would have to be renderable false relative to some context.

  - Now, $\lceil \mathsf{if}\,\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi, \varphi \rceil^c$ tests for whether $i_{c\dotplus \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi} \subseteq \lceil \varphi \rceil^{c\dotplus \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi}$.
  - The context $c\dotplus\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi = \langle w_c, i_c \cap \lceil \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi \rceil^c, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$.
  - The proposition $\lceil \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi \rceil^c = \lceil \mathsf{B}\varphi \rceil^{c/\bigstar}$ tests for whether $i_{c/\bigstar} \subseteq \lceil \varphi \rceil^{c/\bigstar}$.
  - The context $c/\bigstar = \langle w_c, r_c, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$, so that $i_{c/\bigstar} = r_c$.
  - And the proposition $\lceil \varphi \rceil^{c/\bigstar} = \{w : (\exists \Phi, f) \langle w, \langle w, r_c, \langle \Phi, f \rangle \rangle \rangle \in \|\varphi\|\}$.

So the most recent test concerns whether $r_c \subseteq \{w : (\exists\Phi, f)\langle w, \langle w, r_c, \langle\Phi, f\rangle\rangle\rangle \in \|\varphi\|\}$. Suppose this test is passed (we really do really believe φ): that every world in r_c is a φ-world. To keep things simple, let's assume φ is nonepistemic and contingent: P_4 will do. In that case, we know this much: $w_2 \notin r_c$.

And in that case, the context $c + \mathsf{FROM}^{\bigstar}\mathsf{B}\varphi = \langle w_c, i_c, \langle r_c, n_c^1, n_c^2, \ldots\rangle\rangle$; so $i_{c+\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi} = i_c$.

Is the outer test passed? Well, that depends on whether $i_c \subseteq \lceil\varphi\rceil^{c+\mathsf{FROM}^{\bigstar}\mathsf{B}\varphi} = \{w_0, w_1\}$.

Now, i_c is supposed to be the 'discourse-initial' context: free from weird ungrounded suppositions and the like. So it would be pretty odd if it were not the case that $i_c = r_c$. Since $r_c \subseteq \{w_0, w_1\}$, so is i_c. So it would be pretty weird if the outer test were not passed.

I conclude that the infallibility direction does not readily lend itself to the sort of assault to which the omniscience direction yielded. That is not so surprising: familiarly, it is very easy to provide counterexamples to omniscience, but very difficult to provide counterexamples to infallibility. To do so, we would have to admit we are wrong about something! And it certainly never *seems* that way, at least not in the moment.

Note that the argument relies on the identity of i_c and r_c discourse-initially. By contrast, other subjects are not so blessed. Instead of index-shifting to $\bigstar$ I could shift to s_{666}. I can easily imagine Cheney's mistakes.

This may be why, in order to conceive of myself as mistaken, I need to 'alienate' myself from myself: this Hellie fellow has the wildest opinions! Don't trust him. Or: what a fool I was!

- We noted at the outset that the contrapositive of infallibility (if $\neg\varphi, \neg\mathsf{B}\varphi$) is a theorem but the contrapositive of omniscience (if $\neg\mathsf{B}\varphi, \neg\varphi$) is not.

  What this contrast amounts to is the following:

  - For omniscience, the psychology of considering the contrapositive is something like the following.

    We are invited to consider a context c in which we are not certain that P_4.

    If c is one in which we are certain that $\neg P_4$, then, from the perspective of c, we grant: $\neg P_4$.

    But if c is a context in which we are uncertain whether P_4, then from the perspective of c, we do not grant: $\neg P_4$. Of course not: what it is

to be uncertain about P_4 is to be able to find some worlds among the context with respect to which P_4 is the case, and others with respect to which $\neg P_4$ is the case.

Considering the former worlds, we do provisionally become certain that P_4; considering the latter worlds, we provisionally become certain that $\neg P_4$. This pattern of shifting of certainty is possible because $\mathsf{B}\varphi$ and $\neg\mathsf{B}\varphi$ are nondissective.

In light of this shifting certainty, we find no stable resting point of at which our epistemic commitments are compatible with those required by treating c as the main context. This sort of indecision about which world we are in—marked by a characteristic consideration of various incompatible possibilities, each of which we are willing to take seriously—is faithful to the phenomenology of uncertainty.

It may be that we use some aspect of our index shifting operator FROM to fix c and return to it from the provisional subcontexts: we assign c to one of the sequence of random information variables made available by the structure of context. Making this rigorous would be tremendously complicated, so I will leave the point at an informal level.

– The psychology of considering the contrapositive of infallibility is very different. We are asked to temporarily take up the perspective of a context c' in which $\neg P_4$ is a certainty.

  But since P_4 is dissective, there is no context in which our information is more specific than that given in c' in which $\neg P_4$ can be affirmed, even provisionally.

  Regarding $\neg P_4$ as something to take for granted, then, significantly constrains the attitudes we are provisionally willing to take toward P_4 (and therefore toward $\mathsf{B}P_4$), in a way that regarding $\neg\mathsf{B}P_4$ as something to take for granted decidedly does not. Until we abandon our commitment to taking $\neg P_4$ for granted, no provisional commitment compatible with this commitment can result in taking P_4 seriously.

The contrast is especially vivid when the commitment is to 'real' belief rather than the sort of provisional acceptance set off by conditionalization. Recognition of our own ignorance is psychologically much less demanding than recognition of our own error. The latter, unlike the former, requires the breaking of a prior commitment. In general, we admit mistakes only when pushed to the point of contradiction: when the alternative to abandoning the earlier commitment would be defectiveness. By contrast, rendering our picture of the world more specific is the essence of inquiry.

This is reflected in the patterns of perplexity we see in the philosophical tradition: for example, in the theory of perception, while hallucination raises grave difficulties, the world outside our view doesn't. (Exception: the 'phenomenal sorites', made difficult to understand by our ready acceptance of the view that certain 'core' aspects of a perceptual state require and admit no special inspection in order to grasp fully; a similar perplexity associated with ignorance of the sort observed in 'inattentional blindness'.)

Let me qualify the point, though: this is so only when the subject-matter is dissective. There is no pressure toward monotonicity in either psychological self-ascription or normative assessment. (Veltman explains the nonmonotonicity of default judgement—'presumably, φ'—via a more complex appeal to nondissectiveness.) We may perhaps operationalize the notion of the objective in terms of that concerning which ignorance is a dialogically live option and error is not: for which nonmonotonicity requires crash.

Part III
Sensory copular verbs

13 Sensory verbs as copular verbs

13.1 Data

- Target constructions:
  - *o* looks (sounds, smells, feels, tastes) ADJ
  - *o* looked (will look, could look, might look, should look, has looked, will have looked) ADJ
  - *o* looks like [PRED DP]
  - *o* looks similar to [DP]
  - *o* looks as if S
  - *o* looks ADJ to one
- Related good constructions:
  - *o* is (will be, was, could be, might be, should be, has been, will have been) ADJ
  - *o* is like DP
  - *o* is DP
  - *o* is like [PRED DP]
  - *o* is similar to [DP]
  - it is as if S
  - *o* seems (seemed, will seem, could seem, might seem, should seem, was seeming, will be seeming, is seeming, has seemed, will have seemed) ADJ
  - *o* seems to be PRED
  - *o* seems like DP
  - *o* (it) seems as if S
  - it seems that S
  - *o* appears (appeared, will appear, could appear, might appear, should appear, (?) was appearing, (?) will be appearing, (?) is appearing, has appeared, will have appeared) ADJ
  - *o* appears to be PRED
  - *o* (it) appears as if S
  - it appears that S
- Related bad constructions:
  - * *o* looks [non-PRED DP]
  - * *o* looks [PRED DP]
  - * *o* (it) looks that S
  - * *o* sounds (feels, tastes, smells) to be ADJ (but, idiomatically, *o* looks to be ADJ)
  - * *o* is to be ADJ
  - * *o* is [non-PRED DP]
  - * it is that S
  - * *o* appears like DP
  - * *o* appears that S
  - * *o* seems that S
  - ? *o* is ADJ to one

- Explicit marking of aspect:
  - *o* is (was, had been, will (should, might) be (have been)) being ADJ
  - *o* is (was, had been, will (should, might) be (have been)) appearing ADJ
  - *o* is (was, had been, will (should, might) be (have been)) seeming ADJ
  - *o* is (was, had been, will (should, might) be (have been)) looking ADJ

13.2 Hypothesis

This data seems to support the hypothesis that the sensory verbs 'look', 'feel', 'sound', 'taste', and 'smell' are copular verbs. In particular, that they are copular verbs *with mandatory adjectival predicative complements*: no descriptions allowed.

So they are neither raising nor control verbs: they cannot take infinitival VP complements. And they are not attitude verbs: they cannot take CP complements either.

What semantic ramifications does this have?

14 A semantics for sensory verbs

14.1 Data

1. There are five of them. They are sensory. We believe pretheoretically in five senses: not just us, but everyone. This seems to be a human psychological universal.

2. They correspond with what we might call 'subject-oriented' sensory verbs: 'Sam looked at the widget', 'Sam felt the widget', 'Sam smelled the stew', and 'Sam tasted the stew' are obvious; in older use we also see 'Sam sounded the conversation'—a usage that survives in, e.g., *sounding for the depth of the sea-bottom* and *sounding out the opposition's willingness for a deal*.

3. When Sam feels (etc.) the widget, this puts her in a position to render a *demonstrative thought* about the widget. 'This here widget', she thinks, 'is a very fine widget'.

4. No restrictions on the adjectives that can appear as complements: 'this looks expensive'; 'that feels like it was designed by Yabu Pushelberg' (buttery soft, namely).

5. Bloodless disagreement: pointing at a fish in a tank . . .

   — 'it looks like it's straight ahead'

   — 'no, it looks like it's a little off to the left'

   — 'you're wrong!'

   — 'no, you're wrong!'

   — [berserker struggle to the death ensues]

   . . . sillies!

6. Patterns of entailment and nonentailment:

   (a) The *phenomenal principle*: 'NP looks ADJ' entails 'NP is ADJ'. So says Howard Robinson. That's not exactly right. What is true is rather:
   The *conditionalized phenomenal principle*: suppose we are willing to go by looking. Then: 'NP looks ADJ' entails 'NP is ADJ'.
   The phenomenal principle is alluring insofar as we have a great fondness for going by looking.

   (b) The *converse conditionalized phenomenal principle*: (going by looking:) 'NP is ADJ' entails 'NP looks ADJ'.

   (c) The *contraposed conditionalized phenomenal principle*: (going by looking:) 'not: NP is ADJ' entails 'not: NP looks ADJ'.

   (d) The (false) *contraposed converse conditionalized phenomenal principle*: 'not: NP looks ADJ' entails 'not: NP is ADJ'.
   No it doesn't. Looking might fail to reveal all.

   What we observe here is an analogue to our Moore-inferences. The CPP is an infallibility inference, the converse CPP is an omniscience inference; and to the extent that infallibility can't be met with counterexamples, its contraposition is valid; but to the extent that omniscience can, its contraposition isn't.

   Despite this, . . .

7. Use in hedging. 'It *looks* like an original Margaret Keane; but it is *really* a clever forgery'.

14.2 Hypothesis

Attitudes *de hoc*

Philosophy is best at thinking about thoughts with propositional contents: sets of worlds used to represent how from an objective point of view subjects picture things as being. Intentional states characterized with these are sometimes called 'attitudes *de dicto*'.

Recently (well, 45 years ago) we've gotten involved with self-predicative contents: sets of subject-centered worlds used to represent how from the subjective point of view subjects picture themselves as being. Intentional states characterized with these are sometimes called 'attitudes *de se*'.

No reason not to think in terms of thoughts with contents predicative of other things: sets of object-centered worlds used to represent how from the subjective point of view subjects picture externalia as being. We could call intentional states characterized with these 'attitudes *de hoc*' ('about that', if I'm not mistaken: anyone know any Latin?).

Plausibly when we are happily hammering away we are not picturing the world as a whole; nor are we focused narrowly on ourselves; rather we are focused on hammer and nail and events of striking. So perhaps attitudes de hoc are a good model for the engaged perspective.

Perception and the de hoc

The standard of truth of a belief de dicto is the actual world. The standard of truth of a belief de se is oneself (now). What is the standard of truth of a belief de hoc?

Answer: an object of a subject-oriented perceptual verb. When Sam looks at the widget, the widget thereby becomes the standard of truth for de hoc belief.

This is a good answer because the standard of truth is fixed by something 'outside' the thought itself. My beliefs have no constitutive bearing on whether I am in the actual world. My beliefs have no constitutive bearing on whether I exist. My beliefs have no constitutive bearing on whether I am looking at a hammer.

We can think of the availability of five ways a standard of truth might be fixed: by looking, by feeling, etc.

Sentences and the de hoc

Sentences need subjects and have propositions rather than properties as their contents, so de hoc beliefs cannot be expressed in language.

However, we can come close. Looking at a widget, Sam predicates de hoc *well-made*. The belief is true just if the widget she is looking at is represented by

the content of her de hoc belief—namely, just if it is in the class of exactly the well-made possibilia.

Sam can then express something like this belief as follows. We can think of a 'perceptual demonstrative' like 'this' as securing its content in a context from the contextual standard of correctness of de hoc thoughts.

And we can certainly think of having ordinary linguistic predicates share content with beliefs de hoc. Perhaps Sam's predicate 'well-made' shares its content with her de hoc thought.

In that case, Sam can produce a sentence with a de dicto content which is true just if her de hoc belief is: namely, 'this is well-made'.

(I need to say more about this relationship but right now I don't have time.)

Test semantics

The similarity to the Moore-inferences suggests treating sensory copular verbs with a test semantics.

Theory: 'looks well-made' tests a certain subclass of the de hoc beliefs (those with the standard of truth set by looking) for whether they entail 'well-made'. It expresses the necessary property if so; and the impossible property otherwise.

Then 'this looks well-made' performs the same test. It expresses the necessary *proposition* if so; the impossible proposition otherwise.

Accommodating the data

1. If there are five sensory pathways by which the standard of truth can be established, this is surely a psychological universal; if perceptual copular verbs test the contents of the de hoc beliefs based on these pathways, that explains why there are five of them.

2. The object term of such a verb concerns the entity which is the standard of truth set by the appropriate pathway.

3. This falls out of the tie of demonstrative concepts to the standard of truth.

4. The contents tested are *belief* contents and not *sensory* or *perceptual* contents. Belief is responsive to tons of information beyond the immediate environment.

5. Test semantics predicts this. One guy is testing his de hoc beliefs, the other guy is testing *his* de hoc beliefs. Different subject-matter, no disagreement.

6. Test semantics predicts this.

7. This is probably a case of perspective-hopping combined with genericity: our guy is thinking of the typical person's de hoc beliefs about the thing in the former conjunct while expressing his own opinion in the latter conjunct.

14.3 Consequences

- This shows nothing about the contents or metaphysics of perception. The tests are entirely internal to de hoc belief. The language leaves us with no information whatever about how perception influences such belief.
- Nevertheless it is somewhat striking that de hoc belief refuses to categorize with nouns, insisting only on qualifying with adjectives.

Part IV
Action primitivism

15 Basics

Two competing perspectives on the nature of action: *internalism* and *externalism*.

15.1 Internalism

A relatively unquestioned orthodoxy. An action (-kind) (*running*, *walking to the office*, *mentally calculating*) is a physiological process (-kind) (*exhibiting running behavior*, *exhibiting walking-to-the-office behavior*, *'exhibiting' neurological processes of the 'mental calculation' variety*) 'under the governance' of a psychological process (-kind) (*trying to run*, *trying to walk to the office trying to mentally calculate*)—or perhaps a persisting psychological state (-kind) (*intention to run*, *intention to walk to the office*, *intention to mentally calculate*).

Action has a 'conjunctive' character: when one acts, this involves a physiological process being under the governance (whatever that is) of a psychological process.

In a particular instance, the psychological story is therefore complete by the time the trying is in the world. No psychological kind has by necessity any characteristic behavioral manifestations (trying to run though paralyzed, trying to walk to the office though continually beset by salesmen and evangelists, trying to calculate though . . .); and even in cases in which such manifestations are, by stroke of good fortune, manifest, in cases in which the action is success-individuated (running, no; walking to the office, yes), the psychological imposes no conditions on the future.

15.2 Externalism

An action kind is a psychological kind that of necessity requires physiological manifestations; and which, in some cases, imposes constraints on the future.

For example, if one is running, the manifestation of running behavior is a 'part' of the running action. Or if one is walking to the office, one's eventual arrival at the office is a 'part' of the ongoing action of walking to the office.

Trying to run (walk to the office), though a psychological kind, is in some sense 'derivative on' running (walking to the office).

15.3 Contrasts

Suppose that Tweedledee and Tweedledum are both trying to run; but while Dee is succeeding, Dum is not: he is dreaming (paralyzed, etc). Or suppose that Dee and Dum are both trying to walk to the office; but while Dee will get there at some later time, Dum won't: he will pass out and be rushed to the ER.

According to the internalist, in each case, Dee and Dum may well instantiate the same practical-psychological properties.

According to the externalist, in each case, Dee and Dum instantiate different practical-psychological properties.

Is this dispute substantive? It depends on whether the notion of the 'practical-psychological' is substantive.

According to my view, action kinds are kinds of experience, and kinds of experience are fundamental kinds. So according to my view, the dispute is substantive.

15.4 More on externalism

The externalist picture can be fleshed out metaphorically by ascribing a quasi-perceptual aspect to action. To be performing a certain sort of action—to be *G*-ing—is to pick up on, or be 'acquainted with', a certain pathway, and to find one's physiological processes and psychological updates guided along this pathway. When I am walking to work, I find a sort of 'scent trail' concluding in my arrival at work and am guided inexorably along it.

In cases of ignorance and error the manoeuvres are similar to those observed in the philosophy of perception: because I am uncertain of or mistaken about the truth of some proposition on the truth of which which my success rests, I am uncertain of or mistaken about my success; and accordingly about which kind of action I am performing (and accordingly about the character of my stream of consciousness).

15.5 Central questions for the externalist

- Can we be more specific about the notion of 'success resting on the truth of *P*'?

  Yes: this will turn out to implicate the concept of 'knowhow'. Once we see what that is we will be in a position to explain ignorance and error of action in terms of factual ignorance and error.

- What is the rational role of action? If action structures the stream of consciousness and this structure is also a rationalizing structure, action should rationalize or be rationalized or both.

Answer: it does both; the story here is a mix of our story about knowhow and a bit of decision theory.

- What is trying? How can we make specific the notion of trying to G as 'derivative on' the notion of G-ing? How can we capture apparent analytic equivalences between G-ing and trying to G? How is trying involved with self-ignorance and self-error?

  Answer: taking oneself to be trying is a certain sort of epistemic attitude toward the future in light of one's rational commitments; since the epistemic is nonfactual facts about trying do not have sufficient metaphysical oomph to ground facts about success.

15.6 Some data

Some important data about actions:

- Categories of properties:
  - Dynamic properties
    Either:
    - If Fs at t, t is the first and last moment at which Fs ('for the time being', that is)
    - If Fs at t, t is neither the first nor the last moment at which Fs
    - Instantaneous (the former category)
      'Achievements'
      summit Mt Everest, *arrive in Paris*, *finish this book*, *start making dinner*
    - Durative (the latter category)
      - Telic (if s is G-ing at t, it does not follow that, at t, s has yet Ged; there is a sensible and useful sense of 'culminatedness' applying to courses of G-ing from the perspective of times)
        'Accomplishments'
        climb Mt Everest, *travel to Paris*, *write this book*, *make dinner*
      - Atelic (not so)
        'Activities'
        run, *write*, *climb up Mt Everest travel toward Paris*, *cook*
  - Static properties (not so: for F static, there exist possible s and t such that one is so and also such that the other is so)

- The 'imperfective paradox'

  The following story seems coherent:

  > Bill was crossing the street when he dropped dead and never made it across.

  Many linguists have understood this to show that, for G an 'accomplishment' (dynamic durative telic), from s is G-ing at t, it does not follow that for some $t' > t$, at t', s has Ged.

  I think this is a mistake. As I understand it, the role of the progressive is to topicalize a nonfinal subinterval of the interval of time in which the topic event is regarded as taking place. What kind of event is the topic event? And over what interval of time is it regarded as taking place? Indeed, when the progressive is absent, as in 'that rock rolled down the hill', how are we to understand what the topic event is to be? Do we need to know the laws of physics to understand this?

  These questions are much much easier to answer if we regard 'accomplishment predicates' as applying only to culminated events. Since basic linguistic processing is very stupid, we should strongly prefer theories on which questions are very easy to answer.

  This theory also allows us to explain the attractiveness of our little story. Narrowing of attention is a ubiquitous challenge for externalists. Consider the following story about Cheney's heart:

  > That very thing came into existence ex nihilo in outer space.

  If this seems coherent, function and species are not essential properties of organs.

  Here the externalist can say that the story seems coherent only because of what we are leaving out. Filling out data about the topic of discussion results in the following (more plausibly) incoherent story:

  > Cheney is a human so his heart is a human heart. That human heart came into existence ex nihilo in outer space.

  Anything that comes into existence ex nihilo in outer space is not a heart, let alone a human heart.

The question then is what are we imagining when we find the other story coherent? Plausibly that something resembling Cheney's heart came into existence ex nihilo in outer space.

Saying the same about the imperfective paradox raises the question of what we are saying in that relatively commonplace kind of story. Plausibly that Bill was doing something similar to crossing the street when he dropped dead of a heart attack. But similar how? Plausibly from the first-person perspective: for Bill it was as if he was crossing the street—namely, he was trying to cross the street.

Note that for processes that involve no first-person, the imperfective paradox is much harder to get going:

> That rock was rolling to the bottom of the hill when it came to rest in a thick place of grass.

There we find it hard to see in what sense it was rolling to the bottom of the hill. Rolling *down* the hill certainly. But there is no especially obvious variety of resemblance between complete rolls to the bottom of the hill and merely partial rolls that someone telling this story could easily be understood to have in mind.

- A few 'logical' points about actions:
  - Actions are *multiply realizable*: (nearly) anything can get done in any range of ways
  - Actions are *divisible*: for (nearly) any specific way of performing an action, it can be broken into steps that are themselves actions
  - Actions are *conditional*: (nearly) any time we perform an action, the specific way we are to perform it is not fixed in advance, but needs to be adjusted—sometimes massively, sometimes subtly—in response to incoming evidence

These principles suggest a priority of the governing action over its means.

Multiple realizability and divisibility strongly suggest that actions are 'gunky': (nearly) any particular action has a realization of a distinct kind; and (nearly) any such realization is composed of parts which are themselves actions. If these are true without exception and there are no 'minimal actions': actions which are realizers but not realized. Gunkiness demands a 'priority' of governing actions over their means.

Realizability also combines with conditionality to yield this sort of priority (at least a failure of priority in the opposite direction). Suppose a sort of 'open future' view. Grant that when I start *G*-ing, its realization is not determined in advance, but remains indeterminate until the bitter end (its culmination), the governing action determinately exists though its eventual realization does not. And grant that the determinate cannot depend on the indeterminate. Then governing actions do not depend on their means.

16 Knowhow

What is it to know how to *G*? Let's focus on accomplishments: talk of 'knowing how to run/cook/bike/lie on the couch/play guitar' is semantically peculiar (perhaps because activities are metaphysically peculiar).

16.1 Questions, commands, and conditionals

A view on the semantics of speech acts other than assertions:

- The semantic value of a question is surely not a proposition. What is it? Let's say it's a *set of answers*.
- What's an answer? Consider 'who killed Mr Body?' (Suppose it was Prof Plum.)

  In one sense an answer to that is a speech act or maybe a phrase that gives enough for the truth: 'Prof Plum'.

  In another sense it is a sentence that fully expresses what is conveyed in such a case: 'Prof Plum killed Mr Body'.

  In another sense it is the *meaning* of such linguistic entities: Prof Plum himself, or the proposition that Prof Plum killed Mr Body.

  In another sense, truth is not required: just anything sensible. 'Miss Scarlet killed Mr Body' or the proposition that Miss Scarlet killed Mr Body are answers in this sense; 'Berlusconi is a jerk' and the proposition that Berlusconi is a jerk are not.

  Our notion of an answer is *sentential*, *semantic*, and *permissive*: any semantic object that is expressed by a sensible answer to the question is an answer in our sense.
- Not all answers are propositions. It is plausible that 'if Rahm primaries Obama, who will be the GOP nominee?' and 'who will be the GOP nominee?' have the same set of *propositional* answers: every proposition of the

form *the GOP nominee will be X*. But the questions have different meanings in some sense, and we might want our theory to reflect this.

So we should say that an answer to the conditional question is a conditional semantic object: whatever it is that answers to 'if Rahm primaries Obama, the GOP nominee will be *X*': perhaps the pair of the proposition that Rahm primaries Obama and the set of all propositions of form $\langle$*Rahm primaries Obama*, *the GOP nominee will be X*$\rangle$.

- It is not obvious that all semantic objects (of full sentences) are propositions or constructions out of propositions.

  For example, the imperative 'close the window!' may have as its semantic value a property: in particular, the action-kind *closing the window*.

  We might think this because imperatives are entirely tenseless and aspectless:

  * Closes the window!
  * Closed the window!
  * Have closed the window!
  * Has closed the window!
  * Had closed the window!
  * Be closing the window!
  * Have been closing the window!
  * Has been closing the window!
  * Had been closing the window!

  Rendering their semantic values into something richer than the semantic value of a VP would overgenerate.

  To be really confident in this argument we'd need to check conjunction data.

- Imperatives can appear in the consequents of conditionals, as we noted two handouts ago:

  – If it gets cold, close the window!
  – If you're traveling to north country far, where the wind blows heavy on the borderline, remember me to one who lives thar!

- Imperatives can be issued as the answers to questions:

- What do you want me to do?
  Close the window!
- How do I make guacamole?
  Get some avocados! Remove their flesh into a bowl! Mush it up! Chop some tomatoes! Chop some shallots! Stir that in with the avocado flesh! Stir in some lime juice! Season to taste!
- How do I make guacamole if I don't have any avocados?
  Make some scrambled eggs! Put them in a bowl! Mush them up! Chop some tomatoes! ...

Indeed, this seems to be the normal form of answer to a how-to question, as we see from the 'Instructables' web site.

(I did an experiment here, asking 'how do I boil water?' in a FB update and got back imperatives as answers.)

- If all this is right, then perhaps we could represent answers to 'how-to' questions like 'how do I G when P?—we could call such answers *instructions*—as triples $\langle P, G!, M! \rangle$, where P is a proposition representing the informational basis of the question, $G!$ is a property representing the mystery action, and $M!$ a property representing the recommended means.

  (Why just the imperative in the antecedent? Once again regarding it as involving a full proposition seems to overgenerate: the following questions seem to be equivalent:

  - How do I make 'oysters and pearls'? (A famous signature dish of the chef Thomas Keller)
  - How do you make 'oysters and pearls'?
  - How does one make 'oysters and pearls'?
  - How do they make 'oysters and pearls'?
  - How does Thomas Keller make 'oysters and pearls'?
  - How to make 'oysters and pearls'? (this is a bit weekly magazine-ish but works)

  This equivalence suggests that the subject term plays no semantic role. The 'Thomas Keller' case is interesting: it wouldn't seem to be a helpful answer to say 'he doesn't: his cooks do it for him'. But 'how does Charlie Trotter make 'oysters and pearls'?' would seem to be asking something else—perhaps 'how does one make 'oysters and pearls a la Charlie Trotter?'.

  Moreover, note that tense and aspect play no role in how-to questions:

  - How do I get to Carnegie Hall?
  - * How do I got to Carnegie Hall?
  - ? How did I get to Carnegie Hall? (This is a factual question, quite different from a request for instruction.)
  - * How do I have gotten to Carnegie Hall?
  - * How do I be getting to Carnegie Hall?

  The impossibility of introducing tense or aspect markers into a request for instruction suggests that factual information cannot be what we hope to get in answer to such a question.)

- One *expresses* an instruction with a linguistic object of the following form:

  - To G when P, M!

  Such a linguistic object is in some sense equivalent to objects of the following forms:

  - If P and you want to G/you are G-ing, M!
  - To G when P, you should/can/might M.
  - When P, to G, M!
  - G-ing/want to G? If P, M!

- In that case, the semantic values of how-to questions would be triples $\langle P, G!, \mathscr{S} \rangle$, where $\mathscr{S}$ is a set of kinds of action. (Which one? All of them? Those which are in some sense 'plausible' as ways of G-ing when P? I don't know.)

16.2 Pragmatics beyond assertion

- Among the features determining the state of play in a context are:

  - The 'context set', a proposition representing the scope of our collective uncertainty;
  - A list of significant questions;
  - A list of practical matters governing the activities of the 'addressee' of the context.

- For the unconditional:

- An assertion that P updates the context by conditionalizing the context set on P;
- The question $\mathsf{Q}?$ adds its set of permissible answers to the list of significant questions;
- The imperative $M!$ adds the property *M-ing* to the list of practical matters.

- For the conditional:

  - The indicative conditional $\mathsf{if}\, H, P$ tests the context set for whether every H world is a P world;
  - The conditional question $\mathsf{if}\, H (\wedge \mathrm{G}!), \mathsf{Q}?$ adds its set of permissible answers to the list of significant questions;
  - The instruction $\mathsf{if}\, H \wedge G!, M!$ adds the triple $\langle H, G!, M! \rangle$ to the list of practical matters.

  (There is something vaguely unpleasing about the lack of parallelism between the indicative assertion and the others: ideas?)

- Obviously we need some links among these various components:

  - Suppose that $\langle H, G!, M! \rangle$ is on the list of practical matters. And suppose that the context set entails H; and suppose that $G!$ is on the list of practical matters. Then $M!$ is on the list of practical matters (or the context is defective).
  - Not sure what the impact of performing cut on a question is. Ideas?

- We noted a 'sort of equivalence' between an instruction and certain statements of other forms.

  What this amounts to is the following (this approach is in many respects indiscriminable from that advanced by Charlow):

  - A straightforward instruction 'to G when P, $M!$' or 'if you're G-ing when P, $M!$' (or command '$M!$') 'direct injects' the triple $\langle P, G!, M! \rangle$ (or the action kind *M-ing*) onto the list of practical matters;
  - A 'normative' statement of form 'you should M' tests the list of practical matters for the inclusion of the action kind *M-ing*;
  - A 'normative conditional' 'if you're G-ing and P, you should M' tests the list of practical matters for the inclusion of the instruction $\langle P, G!, M! \rangle$.

16.3 Knowledge

A view on the semantics and pragmatics of knowledge ascriptions:

- Among the features determining the state of play in a context are also:
  - For each significant question, markers of: who has an opinion on its answer, whether we are willing to trust that opinion, and (in some cases) what that opinion is.
- 'Sam knows Q' gets a test-semantics: relative to a context c, the test is passed just if, at c: Q is on the list; Sam is marked as having an opinion about it; and this opinion is marked as trusted.
- The role of trust is as follows. Suppose that Q is on the list, and Sam is marked as a trusted source on its answer. Suppose we then mark Sam's opinion as encoding exactly *this* answer to Q. We then must update by conditionalizing the context set on Sam's answer to Q.
- Propositional knowledge ascriptions are posterior in the order of explanation. 'Sam knows that P' gets a test semantics. Context c passes the test just if: *whether* P is on the list; Sam is marked as having an opinion about it; this opinion is marked as trusted; and the opinion is marked as having the 'positive' valence—in other words, Sam's answer is 'yes'.
- Same thing for knowledge and instruction. Suppose that a request for instruction $\langle P, G!, \mathscr{S} \rangle$ is on the list of significant questions, and that Sam's opinion is marked as trusted.

  Then suppose we find out that Sam's opinion is the instruction $\langle P, G!, M! \rangle$.

  In that case, we put the instruction $\langle P, G!, M! \rangle$ on the list of practical matters.

16.4 Knowing how

Putting this all together: our claim is that 'Sam knows how to G when P' tests the context for whether $\langle P, G!, \mathscr{S} \rangle$ is on the list of significant questions, and marks Sam's opinion as trusted.

If we then go ask Sam 'how do I (you, they) (does he, does she, does one) G when P?', and Sam's opinion is revealed to consist of the instruction $\langle P, G!, M! \rangle$, the trust-marker compels us to add that instruction to the list of practical matters.

Contrary to the claims of some, there is no basis for saying that know-how is propositional knowledge; indeed, that doctrine seems to be at odds with straightforward data.

17 Knowhow, consciousness, and self-knowledge

17.1 The philosophy of action perspective

Is there any 'objective basis' for claims of knowhow? Surely some instructions are good and others are bad. Alternatively: is there any attitude we take toward the world in putting an instruction on our list of practical matters?

The question is especially pressing for the externalist: after all, if actions are 'ingredients of the world', then we would expect there to be a matter of fact whether, if one *M*s when *P*, one will thereby have *G*ed.

Here's what I've been able to come up with ...

- Let's switch our semantic apparatus from regular propositions marked by P to centered propositions marked by F.

  And let's identify action-kinds with sets of 'traces', where a trace is something like a centered world with a longlasting center: a triple $\langle w,s,I\rangle$ for I a temporal interval of positive duration.

- We will say that $|G|$ is the set of all traces in which at the world of the trace, the subject of the trace commences *G*ing and eventually culminates that *G*ing exactly over the course of the interval of the trace.

- And we will say that $||G||$ is the set of all centered worlds $\langle w,s,t\rangle$ such that for some I such that $t \in I, \langle w,s,I\rangle \in |G|$.

  Then we could say that in believing oneself to be *G*-ing, one self-ascribes the centered proposition $||G||$.

Presumably a 'good' instruction $\langle F, G!, M!\rangle$ would be one such that if one who is F commences and eventually culminates *M*-ing, one will in doing so thereby commence and culminate *G*-ing. (No requirement that doing so is the only way to culminate the *G*-ing.) How to capture this idea formally?

- Let's say that $|A_F|$ is the subset of $|A|$ consisting of just those traces with initial 'slices' in $||F||$ (the set of all possible individual slices which are F, the centered proposition self-ascribed in self-ascribing F-ness). Define $||A_F||$ in the obvious way.

- Then $\langle F, G!, M!\rangle$ is a good instruction just if $|M_F| \subseteq |G_F|$.

That's a sort of 'condition of formal adequacy' rather than a fully articulated theory.

17.2 Game theory and action theory

So we should ask: What would M have to be like in order to satisfy this condition? As we saw in the initial section, if M is to realize G and M is 'available from the outset' as an instruction, M had better have considerable stability under increasing information; and that means that M had better be conditional and sequential.

We could think of M as an approach to playing a certain game. The game is played by me versus the world: first I do this; then the world reveals that that is so by giving me evidence. I countermove by doing something else; then the world reveals that something else is so. And so forth. We can think of me 'winning' this game if I culminate a G-ing before death (or before some fixed time); 'losing' otherwise.

The game theory guys have a concept to express what we are looking for: the concept of a 'winning strategy'. A winning strategy is, intuitively, a way of playing a game such that no matter what the opponent does, I win. Cashing this out formally is a bit more of a nuisance: here's what I've come up with ...

- Let's say that a *strategy* is a function S accepting as argument a finite sequence of command-proposition pairs $\langle A!, F\rangle$, and returning a command, such that if $\{\langle A_1!, F_1\rangle, \ldots, \langle A_{n-1}!, F_{n-1}\rangle, \langle A_n!, F_n\rangle\}$ is in the domain of S, $S(\{\langle A_1!, F_1\rangle, ..., \langle A_{n-1}!, F_{n-1}\rangle\}) = A_n!$.

- A *winning strategy* is a such a function such that for any sequence $\{F_1, F_2, \ldots\}$, the sequence $\{\langle S\varnothing, F_1\rangle, \langle S\{\langle S\varnothing, F_1\rangle\}, F_2\rangle, \ldots\}$ is 'good'.

Let's bring this back to the domain of application.

- A strategy *relative to* F is a function S_F accepting as argument a proposition of form $F \cap (E_1 \cap \ldots \cap E_n) \cap (||A_1|| \cap \ldots \cap ||A_n||)$, and returning an imperative $A!$, such that if $F \cap (E_1 \cap \ldots \cap E_n) \cap (||A_1|| \cap \ldots \cap ||A_n||)$ is in the domain, $S_F(F \cap (E_1 \cap \ldots \cap E_{n-1}) \cap (||A_1|| \cap \ldots \cap ||A_{n-1}||)) = A_n!$.

- Restrictions: $F \cap (E_1 \cap \ldots \cap E_n) \cap (||A_1|| \cap \ldots \cap ||A_n||)$ is nonvacuous. [***mean same thing by barbar here?]

- A strategy *for Ging when* F is a such a function such that for any sequence $\{E_1, E_2, \ldots\}$ terminating at the subject's death (or some earlier time), the proposition $F \cap (E_1 \cap E_2 \cap \ldots) \cap (||S_F\varnothing|| \cap ||S_F\langle S_F\varnothing, E_1\rangle|| \cap \ldots)$ entails that the subject has Ged.

- We appeal to a notion of 'grasp' of a strategy for Ging when F as, for the time being, primitive: to operationalize, we could say that such grasp is revealed in production of it as a candidate instruction of how to G when F.

- Say that one *has the skillz* to carry out a strategy S_F just when, for any finite sequence $\{E_1, \ldots, E_n\}$, if $S_F(F \cap (E_1 \cap \ldots \cap E_{n-1}) \cap (||A_1|| \cap \ldots \cap ||A_{n-1}||)) = A_n!$, one knows how to A_n when $F \cap (E_1 \cap \ldots \cap E_n) \cap (||A_1|| \cap \ldots \cap ||A_n||)$.

- Then, we can say that

  - One knows how to G when F just if:
    for some strategy for Ging when F, one grasps it and has the skillz to carry it out.

Obviously that's circular. But: (i) maybe it's merely recursive, and always grounds out, if eg 'do nothing additional' always shows up and we treat doing nothing additional under any circumstances as part of a universal skillz-set; (ii) this is compatible with gunkiness, which, as we have seen, is something we might *hope* to predict.

17.3 Following a strategy

We want a notion of *following* a strategy: undertaking the actions it prescribes in light of one's picture of the world. One's following of S makes a certain course of actions intelligible just if one performs that course of actions out of following S only if that course of actions is prescribed by S in light of one's evolving beliefs.

That isn't a necessary condition on what it is to be following S; rather, it is a limitation on when rationalizing explanations that advert to 'following-S' talk are permissible. Strategies are not metaphysically prior to actions; rather, talk of strategies is a somewhat more precise form of talk about unfolding actions. The need for such talk lies in the multiply realizable character of actions together with the limitations on any individual's access to any given action type.

The dialectic is this. We might say that one is chopping garlic because one is making tomato sauce despite the lack of universality of the explanation: some recipes for tomato sauce involve *slicing* garlic. We might then wonder: why *chopping* rather than *slicing*, in this case?

Here we appeal to a difference in strategies. But the strategy doesn't motivate or rationalize anything: rather, the rational oomph emanates from the act of making tomato sauce. Different strategies are merely different styles of channeling rational oomph into means.

17.4 The link between knowhow and action

One is Ging just if ...

1. For some current *G*ing *g* and some course of actions *a*, one is performing *g* *by* performing *a*, just if

2. For some current *G*ing *g* and some course of actions *a*, one's performing *g* (in light of one's beliefs) *provides the rational motivation for* one's performing *a*, just if

3. For some current *G*ing *g* commenced when *F* and some course of actions *a* one performs just when one performs *g*, one performs *a in the course of following one's best strategy* (the best strategy that one grasps and for which one has the skillz) for *G*ing when *F*—just if

4. For some *F*, one is *exercising knowledge how* to *G* when *F*.

Here we see a chain of equivalences between action, the by-relation, practical rationality, following of strategies, and knowhow.

Justification for the equivalences:

1. The first equivalence follows from the structural points about action with which we commenced;

2. The next is the Michael Thompsonesque doctrine that action is rationalized by action (that what makes one reasonable in commencing a certain action is that it is a means to some unfolding governing action);

3. The next is an explanatory hypothesis about that in which the 'by'-relation consists;

4. The next is an explanatory hypothesis about that in which the exercise of knowhow consists.

17.5 Learning how as growing acquaintance

Our view is that what it is for an update to be reasonable is for the stream of consciousness to require the update in order to sustain coherence. If so, then since (in light of the Thompsonesque doctrine) which actions one is performing make a difference to which updates are reasonable, the kinds of actions one is performing qualify one's stream of consciousness.

But they do so only partially, since the fine-grained motivational story appeals to *strategy* rather than to full-bore action, and we have declared that this shows strategy to be a sort of 'aspect' of action. For this reason, it is only aspects of the kinds of actions one is performing that qualify the stream of consciousness.

This sheds light on what it is to grasp a strategy. The traditional theory of acquaintance makes for a distinction between 'active' and 'latent' acquaintance: one is 'actively' acquainted with the kind (or kinds) of experience one is currently undergoing (or imagining); latently acquainted with the kinds one in some sense 'can' imagine (where latent acquaintance typically follows on active acquaintance but does not precede it).

Now, if the kinds of experiences are the kinds of actions, we can provide further detail to this theory: one is actively acquainted with a kind of action just if one is performing it (in reality or in a simulation), while one is latently acquainted with a kind of action just if one knows how to perform it. Or, incorporating the complexity of strategies: one is actively acquainted with an aspect of an action just if one is following the corresponding strategy for performing it (in reality or in a simulation), while one is latently acquainted with that aspect just if one grasps the corresponding strategy.

So we can say the following:

- Grasp of the strategy $\langle G!, F, M! \rangle$ is latent acquaintance with the aspect of G-ing that presents when F via M-ing.

Accordingly, learning how to perform a certain kind of action—learning to perform it under different circumstances, or in different ways—is extending the scope of one's acquaintance with that kind.

17.6 Self-knowledge, -ignorance, and -error

Sometimes I try and succeed. Other times I try and fail. Sometimes I think I am going to succeed and am right; other times I am wrong. I know that whenever I am doing something, I am trying to do it (but not vice versa); and that when I am trying to do something I think I might (and sometimes *will*) succeed.

I want to propose the following theory of self-knowledge and its less attractive cousins, based on these platitudes:

- 'I am trying to G' gets a test semantics: I pass the test just if I am exercising knowledge how to G when F and I am not certain that $\neg F$ (I was not F when I started out);

  and that just if for some M where M *is* a strategy for Ging when F, M is on my list of practical matters.

  (*Is*—not 'it's my opinion that'. More on this shortly.)

- In this way self-knowledge of trying is compatible with but not entailing of certainty of success.

- Certainty of success then comes about through self-knowledge of trying together with certainty about the fact that when I started out, I was F.

- When I am self-ignorant, this comes about by way of uncertainty whether the article of knowhow I exercise is in its conditions for success;

  When I am in self-error, this comes about by way of error about the facts.

- Finally, what about error about the efficacy of a strategy I grasp?

  One might be following a crap strategy, after all: 'here's how you bake a lemon cake: first, get a fish; second, bone it; third, nail the filets to your door frame. you're done!' or: 'to raise a happy well-trained puppy: shout at it regularly and do not exchange affection'.

  This is where the externalist's sense that one is (at least partly) *acquainted* with the kinds of actions one can perform comes in. Since such an approach could not work under any circumstances, one does not have a sufficient conception of the action at issue to be said to have a strategy for performing it. One is out of touch with nature's flows of agentive possibility, under a delusion of grasp.

  For conditionalized strategies: one has a mistaken conception of what it takes to G when F (e.g., a prejudice against certain breeds of puppy). Here one's acquaintance is insufficiently rich: one loses sight of the action under one's circumstances. One might do one's best to ape the outcomes and one might get lucky but that would not be Ging.

  Uncertainty? can't happen, you either have the grasp or lack it.

17.7 Externalist interpretation of the formal approach

Externalism imposes constraints on the formal pragmatics:

- The list of practical matters consists of (a) actions underway; (b) 'genuine' instructions: instructions that reflect partial acquaintance with the target action.

- The pragmatics of acts of instruction (understood as linguistic episodes) therefore takes on a set of commitments resembling those familiar from externalist approaches in other domains: we make a distinction between acts of genuine instruction and acts of 'pseudo-instruction'. If one's instructional attempt is couched as an attempt to expand the audience's range of acquaintance with the target action, but it does not in fact do so, the formalism won't be by itself useful in interpreting the situation: it can instead be understood

as supplying something like an ideal, so that the purport of conformity to it provides clues to interpretation. (Compare ‘mock de hoc thoughts’ and assertions with the absurd proposition as content.)

Part V
Time

18 Fixity and openness

Perhaps some facts are 'determinate' or 'fixed', others 'indeterminate' or 'open'. Perhaps not. This seems like a metaphysical question. (A realm of facts serving as an especially promising candidate for openness, in my view, is those concerning the future decisions of free agents: will the admirals decide to wage a sea battle tomorrow? Perhaps it is not yet fixed.)

18.1 The Williamson-Barnett challenge

Unfortunately, difficulties of expression and interpretation have led to significant logical worries about the very coherence of the doctrine of openness. How shall we express the doctrine that it is open whether φ? The rough idea is that there is no fact that φ and there is no fact that $\neg\varphi$. But the obscurity of the quantification over facts here suggests at least as an initial manoeuvre the replacement of this talk with a combination of sentential operators. I call following dialectic, a challenge to the most immediately obvious proposals, the 'Williamson-Barnett challenge'.

First try:

- $\neg(\varphi \vee \neg\varphi)$

Problem: this reduces by DeMorganization to $\neg\varphi \wedge \neg\neg\varphi$: even without double-negation elimination, already a contradiction.

Second try (where T abbreviates 'it is true that'):

- $\neg\mathsf{T}\varphi \wedge \neg\mathsf{T}\neg\varphi$

Problem: the T-schema

$$\varphi \iff \mathsf{T}\varphi$$

reduces to the two directions

$$\text{if } \varphi, \mathsf{T}\varphi \qquad\qquad \text{if } \mathsf{T}\varphi, \varphi;$$

which contrapose to

$$\text{if } \neg\mathsf{T}\varphi, \neg\varphi \qquad\qquad \text{if } \neg\varphi, \neg\mathsf{T}\varphi;$$

and the former, together with our second try, entails $\neg\varphi \wedge \neg\neg\varphi$.

Third try (where F abbreviate 'it is fixed that'):

- $\neg\mathsf{F}\varphi \wedge \neg\mathsf{F}\neg\varphi$

Problem: the 'F-schema'

$$\mathsf{T}\varphi \iff \mathsf{F}\varphi$$

is independently plausible—surely, if it is fixed that φ, it is true that φ; and if it is true that φ, the truth of φ suffices to fix that φ. The F-schema reduces to the two directions

$$\mathsf{if}\ \mathsf{T}\varphi, \mathsf{F}\varphi \qquad\qquad \mathsf{if}\ \mathsf{F}\varphi, \mathsf{T}\varphi;$$

which contrapose to

$$\mathsf{if}\ \neg\mathsf{F}\varphi, \neg\mathsf{T}\varphi \qquad\qquad \mathsf{if}\ \neg\mathsf{T}\varphi, \neg\mathsf{F}\varphi;$$

and the former, together with our third try, reduces to our second try.

18.2 Against contraposition

The validity of contraposition is crucial in the case against the third try. After all, the F-schema is defended in its 'straight' or unnegated form, but the contraposed form is the one employed in the reductio.

This is problematic. As we have seen, in an update framework, contraposition is not valid:

$$\mathsf{if}\ \varphi, \mathsf{B}\varphi \quad \not\vdash \quad \mathsf{if}\ \neg\mathsf{B}\varphi, \neg\varphi.$$

This is the central datum motivating the update treatment of B: while acceptance requires acceptance of acceptance, acceptance of failure of acceptance does not require rejection.

The reason for this is the availability in the update framework of a class of sentence, the proposition expressed by a member of which depends on the characteristics of the context; together with the Ramsey-test verifying character of the update conditional: the conditional tests a hypothetically updated context for the entailment of the consequent. Facts about set-complementation secure contraposition for conditionals with with contextually-invariant consequent propositions [***true for the contrapositive of the contrapositive?] while failing to do so for conditionals with contextually sensitive consequent propositions.

This suggests the obvious strategy for evasion of the Williamson-Barnett difficulty: interpret F as a sensitive operator. How?

18.3 Fixity and questions

Note to begin with that a natural expression of a fixity claim involves an embedded question: it is natural to say 'it is not fixed whether φ': perhaps even more natural than to say 'it is not fixed that φ'. Not so for truth: 'it is true whether φ' is not easily interpreted, or at least extremely stilted. More generally, it is easy to process 'it is fixed who will fire the first shot in the sea battle'; by contrast, 'it is true who will fire the first shot in the sea battle' is barely intelligible.

This suggests regarding the fixity operator as in the first instance an operator on questions and in the second instance an operator on their answers (compare our treatment of the knowledge operator). In that case the logical form of 'it is fixed whether φ' or 'it is fixed who will fire the first shot in the sea-battle' would be the following:

- $\mathsf{F}(\mathsf{Q}x : \varphi(x))$.

Next issue: what is the meaning of this operator? What attitude are we expressing toward a question when we say it has a fixed answer?

Recall that the semantic value of a theoretical question is a partition of the modal base; the pragmatic effect of the speech act of asking a certain question is the introduction of the partition that is its semantic value to the list of live theoretical issues. So, in line with the general update approach, the fixity operator should function semantically to perform a certain test on the list of live theoretical issues.

I want to suggest that the test in question concerns the context's attitude toward the *answerability* of the question: roughly, to regard it as unfixed whether φ is to regard φ as without an answer; where our attitude toward this absence should be distinguished from a pair of other ways to regard an answer as absent.

At one extreme, we might regard whether φ as without an answer because the question makes no sense. This is not the attitude we take in regarding whether φ as unfixed. A claim of fixity makes sense only if the question embedded makes sense: fixity claims test the attitude of the context toward a certain well-defined semantic value.

At the other extreme, we might regard whether φ as without an answer as a merely contingent matter: as it happens, no one is in possession of the answer. Nonfixity imposes a more stringent constraint. Regarding whether φ as simply unanswered is something like an empirical attitude toward a contingent truth; regarding whether φ as unfixed is regarding it as unanswered not merely as a contingent or empirically assessable matter but out of something a bit more like logic. To regard it as unfixed whether φ is, roughly, incompatible with regarding it as *possible* for anyone to have an answer to whether φ.

We can approach the notion of fixity from another direction. We can make sense of the idea that one question entails another: Q entails Q' just if $\|Q\| \supseteq \|Q'\|$; just if the partition associated with Q is strictly finer than that associated with Q'. Example: 'what did Nixon know and when did he know it' entails 'what did Nixon know'; in answering the former we thereby answer the latter.

Perhaps every context involves a maximally strong question as a live issue: a question entailing every other question the context could coherently regard as a live issue—something along the lines of the question 'what is the case'. Whatever the case may be about any other matter, that is part of what is the case simpliciter. So in answering the maximally general question, every other answerable question is thereby answered.

For an example of these two approaches, consider the question whether there will be a sea battle tomorrow. If we think this question is unanswerable in the intended sense, we think of anyone's claim to possession of an answer as mistaken, and we think of neither answer as part of what is the case: accordingly, pursuit of the answer to the question is idle (we might wait around until the answer becomes fixed and then pursue it, but in doing so we would only be coherent if we changed our attitude toward the question's answerability).

Putting these ideas together, we could say that $\mathsf{F}(\mathsf{Q}x : \varphi(x))$ tests the context for whether the maximally specific question of the context entails $\mathsf{Q}x : \varphi(x)$. If not, we both reject any attempt to add the question to the list of live issues and any attempt to introduce a knowledge claim regarding the question. Finally, we could think of $\mathsf{F}\varphi$ as equivalent to $\mathsf{F}\varphi? \wedge \varphi$.

Resolving the Williamson-Barnett difficulty at this point is a routine matter. if $\varphi, (\mathsf{F}\varphi? \wedge \varphi)$ is valid: updating on φ is answering $\varphi?$ in the affirmative, and therefore regarding the question as answerable. But the contrapositive if $\neg(\mathsf{F}\varphi? \wedge \varphi), \neg\varphi$ is not valid: updating on $\neg(\mathsf{F}\varphi? \wedge \varphi) = \neg\mathsf{F}\varphi? \vee \neg\varphi$ is updating either on $\top$ just if $\varphi?$ is regarded as unanswerable, or on $\neg\varphi$ just if $\varphi?$ is regarded as answerable; in the former sort of context, the modal base will contain a mix of φ and $\neg\varphi$ worlds; and in that case the updated context will fail to entail $\neg\varphi$.

19 Credence and dynamism

19.1 The problem

The semanticist's basic tool is the set of possible worlds. After all, representation is 'as of the world': to depict us as representing anything other than at the very least a *possibility* for the world would be to distort the nature of the enterprise; and representation is 'partial': we should accommodate ignorance through indeterminacy.

That argument leaves something important about representation out: possible worlds have no 'perspective' built in (or at best a 'god's eye' perspective), but representation is always or at least often from a perspective 'within' the world. In particular, representation is often in time and of time: we represent bits of the world as past, others present, others future. Characterizing this aspect of representation requires some fancy footwork.

The standard approach to this task departs from the basics by complicating the base-level ontology: what is represented is characterized by a set of 'centered' possible worlds, where a centered world is possible individual-slice. The effect is something like that of de hoc content where the standard of truth is not a persisting external object but an instant in one's own life.

This is problematic on its face.

- Since the standard of truth is never around for more than an instant, it is not immediately clear how to represent diachronically assembling a body of information about the standard of truth. Certainly the tradition—on which this involves narrowing the proposition or property one ascribes to the standard of truth—is unavailable.

- Since the standard of truth is never around for more than an instant, it is not immediately clear what the bodies of information acquired about different standards of truth have to do with one another.

  Suppose for purposes of comparison that for a while I am hammering, so that my de hoc content concerns the hammer. Then I go play video games for a while, so my de hoc content concerns my avatar. There is, as yet, almost no logical connection whatever between these bodies of content. I could perhaps assemble them into my picture of the world through the use of 'de re' propositional content and some sense of how each object fits into a certain 'map', but I may regard this as unnecessary.

  By contrast, we would imagine that the synthesizing the information we build about ourselves over time is of the highest significance, and that there ought to be a well-regulated logic of how to do so.

(These are problems for the temporal aspect of content de se et nunc; we save the difficulties for its personal aspect for later.)

A well-known, less abstract way of making the point is to think of a clock-watcher: they gain no objective information, but somehow their perspective changes. If we model their perspective using centered worlds, then once they become certain it's 2PM, every centered world is a 2PM world; and that's something that won't go

away as they acquire more information—at least not according to the traditional conception of learning as taking the intersection.

These problems stem in each case from the complication of the modal base. The perspectival aspects of representation shouldn't be inflated into ingredients of the ontology. We need to start fresh, separating perspective and ontology from the ground floor.

But how? Don't we need to pack everything relevant to representation into the modal base? As we know well by now, we certainly do not: update semantics allows us to segregate the representational aspects of thought and communication from their nonrepresentational aspects.

19.2 The open future

Let's ease into my approach by thinking about the open future view in metaphysics. The picture there is botanical. We begin with a big bushy shrub: at the creation of the world, there remain a vast array of ways we might eventually get to the destruction of the world. As time passes, what does in fact happen becomes 'fixed': as possible futures become *merely* possible pasts, the 'trunk' of the shrub gets trimmed of low-down foliage, revealing something more like a tree. Going on, more and more branches are pruned away, leaving fewer and fewer paths to the destruction of the world. At the final moment, only a single possible end-state remains, and the tangled shrub with which we started has finally been denuded into a lifeless pole.

The tree metaphor can be turned into a space of possible worlds picture easily enough. Let's grain up our notion of a possible world: a possible world is now a temporally ordered possible sequence of stages—a sort of total possible history of the world, a total possible way of getting from creation to destruction.

Thinking in these terms, we can locate worlds that share initial segments: w and w' do so just if there is some initial interval I such that at each time in I, w and w' are then each going through the same kind of stage. We can then think of a class of relations $\mathscr{I}_I \subseteq W \times W$, where $\langle w, w' \rangle \in \mathscr{I}_I$ just if at each time in I, w and w' are then each going through the same kind of stage. Each such $\mathscr{I}_I$ will be an equivalence relation, establishing a partition over W, where each world occupies exactly one cell in the partition, and if w is in a certain cell then that cell contains exactly the worlds that are exactly the same as w throughout I (and perhaps further into the future).

We then sharpen our metaphor by thinking of the possible futures of the world at the creation of the universe as containing all of W, which is the unique cell of the partition $\mathscr{I}_\varnothing$. If $I \subset I'$ (so that the end of I' is later than the end of I), the partition induced by $\mathscr{I}_I$ will grain strictly more coarsely than that induced by $\mathscr{I}_{I'}$:

any worlds which still share an initial segment did so earlier as well. In this sense, as time passes, things become more determinate.

This is reflected in our thinking about the standard of truth in a context no longer as tied to the unique world of utterance of the context for all times, such that an assertion of φ at c (where $w_c = w$) is true just if $w \in \lceil \varphi \rceil^c$. Instead, we think of the standard of truth as a cell in a partition induced by one of the $\mathscr{I}_I$: if the time of the context $t_c = \sup I$, so the cell of the context $C_c \in p_c = \mathscr{I}_I$; and an assertion of φ at c as true just if $C_c \subseteq \lceil \varphi \rceil^c$. (Here we could worry about context of assessment stuff, perhaps; but not now.)

19.3 Representing the present

On this picture, the present is the moment at which 'genuine' possibility for the future starts: the past is 'fixed', the future 'open'; the genuine possibilities for the future are those found in the current actual cell.

I don't know if this is how truth and possibility work but it is certainly the way our representation of truth and possibility works. We picture the past as fixed and the future as open. Our theory of temporal representation should reflect this.

Of course, we don't want to tie the possible worlds used in semantics too closely to genuine possibilities for the future: we are uncertain about the past as well as the future. Still, we can recognize an asymmetry in the way we regard these sorts of uncertainty.

The picture is this: think of a forest. I am certain one of the trees is (currently) actual; but I am not certain which one it is. Searching for the truth is zeroing in on our (current) actual tree; accordingly, the nonspecificity in temporal representation can be thought of as indeterminacy in the representation of which region of the forest we are in; propositions can be thought of as regions of the forest.

How to translate this into possible worlds talk? My suggestion is this: we think of the information state of a context not as exhausted by its modal base, understood as a set of possible histories of the world, but as involving in addition a sort of 'overlay' on the modal base. This 'overlay' is in the business of grouping possible histories into what are regarded as trees. Uncertainty about the past is then uncertainty about which ostensible tree is actual; uncertainty about the future can in addition be uncertainty about how the actual tree will be pruned down. This distinction in how possible histories are grouped together reflects an representational asymmetry in regard to the past and the future; one's sense for when the present is can be recovered from the fact that we see the present as where the trunk starts branching outward.

Next, syntax. I'm inclined to think that temporal discourse in English has a logical form something like the following:

- An assertion, modifying the modal base:
  - $R(I_T, I_X)$; and
- A presupposition, modifying the partition:
  - $\mathsf{pres}\,\mathsf{ha}-[I_T, I_R]$;
  - $\mathsf{past}\,\mathsf{ha}-[I_T, I_R]$;
  - $\mathsf{pres}\,\mathsf{ha}+[I_T, I_R]$;
  - $\mathsf{past}\,\mathsf{ha}+[I_T, I_R]$.

I_T is the 'topic interval', I_X is some other interval contributed either by lexical material or by context, and I_R is what we will call the 'reference interval'. R is some relation between intervals; the operators pres and so forth will be explained shortly.

For example:

- Max drives to California during Helga's drive to Oregon
- Max drove to California during Helga's drive to Oregon
- Max has driven to California during Helga's drive to Oregon
- Max had driven to California during Helga's drive to Oregon

Here, intuitively, the information conveyed about the modal base is the same throughout: from the god's eye view, the interval of time occupied by Max's drive is strictly surrounded by the interval occupied by Helga's drive. We can implement that semantically by setting I_T := the interval occupied by Max's drive; I_X := the interval occupied by Helga's drive; and R := the relation of strict surround.

But the presuppositions are rather different:

- The first suggests that the speech time is during Max's drive;
- The second suggests that the speech time is after Max's drive (with no suggestion about Helga's drive);
- The third suggests that the speech time is after the end of Max's drive and before the end of Helga's drive;
- The fourth suggests that the speech time is after the end of both.

We can implement these presuppositions in the following way:

- Let ha$-[I_T, I_R]$ set $I_R = I_T$;
- Let ha$+[I_T, I_R]$ set I_R so that its end is strictly preceded by the end of I_T;
- Let pres locate the present moment strictly before the end of I_R but no later than the end of I_T (if these are distinct);
- Let past locate the present moment strictly after the end of I_R.

Now we can advance test clauses. Our 'overlay' will be a partition on the modal base of the context: an equivalence relation τ_c on i_c, such that at every cell in τ_c there is some interval of time such that every member of the cell is exactly the same throughout that interval. Then we think of each cell as a tree that, by the lights of c, we might be in. Accordingly, if at a cell C of τ_c, I^C is the longest interval of exact sameness, then $\sup I^C$ is a candidate for when the present might be, by the lights of c.

Accordingly, the conditions tested for are the following. For each cell C in τ_c ...

- pres ha$-[I_T, I_R]$ tests whether
  $\sup I^C < \sup I_R\ (= \sup I_T)$;
- past ha$-[I_T, I_R]$ tests whether
  $\sup I_R\ (= \sup I_T) < \sup I^C$;
- pres ha$+[I_T, I_R]$ tests whether
  $\sup I_T < \sup I^C < \sup I_R$;
- past ha$+[I_T, I_R]$ tests whether
  $\sup I_T < \sup I_R < \sup I^C$.

We get the desired predictions if we think of I_R as set to I_X in the third and fourth cases. (That won't be a universal result: consider 'Max had driven to California some time after Helga's drive to Oregon'. Rather, context will usually make clear what is intended for I_R.)

19.4 Comments

- We can get the clock-watchers by assignment of suitable logical forms to sentences like 'it's now noon'.

  For an analogue that gives proof of concept, consider the discourse

- 'Max drives to CA during Helga's drive to OR' ...
- 'Max has driven to CA during Helga's drive to OR' ...
- 'Max had driven to CA during Helga's drive to OR'.

There it seems we acquire no new objective information about which world we are in, but we do represent ourselves as successively later and later in the course of events we regard the world as containing.

- Uncertainty about when it is:

  Is it 1PM yet? Has Max driven to CA yet? These questions are hard to discriminate. (To bring them closer, think of the hour hand on my clock as Max and the 1PM marker as CA.)

  To treat the latter, set I_T := the interval of Max's drive. I say that when we are uncertain about the latter, τ_c has both a cell C where at some world in C, $\sup I_T < sup I^C$ and a cell C' where at some world in C', $\sup I^{C'} < \sup I_T$.

  Uncertainty about when it is is just garden variety uncertainty, enhanced with certain facts about the partition.

  This suggests that, as a motivator for content de nunc, the sort of argument Lewis uses to motivate content de se can't work. (He imagined a failure of de se certainty to supervene on certainty about which world one is in: two omniscient gods, one on the highest mountain and one on the coldest mountain, each confused about which god he is.) Fortunately, he doesn't provide one. (But we might assist: suppose that the world is of two-way eternal recurrence. Am I in this epoch or the next, I wonder. But this is not a very whole-hearted assistance. Actually I don't wonder this, because I believe in the open future, with which the 'future necessity' of continuation of the recurrence hitherto is incompatible.)

- Notice that there is no 'absolute right or wrong' about what is happening *now*: the best I can do is to commit to labeling a certain class of momentary events in the history of the world as I see it as those which have just become determinate, and attempt to characterize what goes on before and after it as determinately as possible.

  This is attractive. Is skepticism about when it is even coherent? We might be wrong about what conventional systems of calendar dating say in relation to more global and tightly stitched aspects of the flow of events. But suppose that I am mistaken about no aspect of the flow of events, and I take myself to be located here in 2011: might it still be that I am in fact in 1648? The mind reels!

20 Decision under uncertainty

Suppose I don't know whether conditions are right for my barrel roll. If so, I get glory. If not, I could wind up with hurt feelings—or worse! What to do?

'Yes', 'no', and 'maybe' cease to cut it at this point. Rational agents turn to *credence* or *subjective probability* together with *cardinal value* and various *optimality models* to establish *preference orders*.

20.1 Credence

Up to now we've been representing information with our 'modal base': a set of possible worlds. We get to keep the modal base, but we add something to it: the *probability distribution!* $C : 2^W \mapsto [0,1]$ is a probability distribution just if:

- $C(W) = 1$;
- If $P \cap Q = \varnothing$; then $C(P \vee Q) = C(P) + C(Q)$.

These are the 'Kolmogorov axioms'.

Credence is a more fine-grained approach to belief. Belief and denial go with extremal credences (1 and 0); uncertainty goes with nonextremal or intermediate credences (values less than 1 but greater than 0). The relationship is analogous to that between quantification and percentages. Universality and vacuity go with extremal percentages (100% and 0%); proper partiality goes with nonextremal or intermediate percentages (less than 100 but greater than 0).

(I should highlight that a number of theorists fail to accept this relatively obvious story: mostly because they are unfamiliar with it, partly because they have been taken in by some alluring but misleading philosophy.)

20.2 Test semantics for probability statements

If we think of i_c henceforth as a credence function rather than a proposition, we can make the following proposals about belief-avowals and statements of probability:

- $\mathsf{B}\varphi$ tests whether $i_c(\varphi) = 1$;

  (excuse the sloppiness about use and mention, please)
- $C(\varphi) = x$ tests whether $i_c(\varphi) = x$.

I will sometimes write $|i_c|$ as a name for the strongest proposition P such that $i_c(P) = 1$.

20.3 Conditional probability

We can think of a two-place 'conditional probability' function $C(\varphi/\psi)$ as induced by the probability function C, via the following rule:

- $C(\varphi/\psi) = C(\varphi \wedge \psi)/C(\psi)$.

Memorize this. Basically, it equates the probability of φ *given that* ψ with the probability of both, renormalized to the probability of the latter. Since this is in fact the probability we would be looking at for φ, if we became certain about ψ (and nothing else changed), that is a nice rule for conditional probability.

We have seen that the update 'if' tests for the containment of the antecedent-worlds within the consequent-worlds, and we can do something similar for claims of conditional probability:

- if ψ, φ tests whether $i_c(\varphi/\psi) = 1$;
- $C(\varphi/\psi) = x$ tests whether $i_c(\varphi/\psi) = x$.

20.4 Value functions

A value function V (not a 'valuation function' of the sort we saw back in the logic stuff—and we're stealing its letter) assigns real numbers to representata on the basis of how awesome they are. The representata in question will be allowed to vary adventitiously.

20.5 Expected value

Expected value is something like a credence-weighted average of value.

- Very simple example: 'how awesome would it be if pigs learn to fly?' Well let's see. Let $V : W \mapsto \mathbb{R}$ be our assessment of exactly how awesome a given world is; let P := the proposition that pigs learn to fly. The answer to the question is this:

  $$\sum_W V(w) \cdot i_c(w/P).$$

  There we took every maximally specific way things might go, assigned it the appropriate degree of awesomeness, and then took a credence-conditional-on-P-weighted average value of those awesomeness levels.

- Getting a bit more germane: 'how awesome would it be if I bring red wine to the party?' Let $R :=$ the proposition that I bring red wine to the party. The answer to the question is this:

$$\sum_W V(w) \cdot i_c(w/R).$$

There we took every maximally specific way things might go, assigned it the appropriate degree of awesomeness, and then took a credence-conditional-on-R-weighted average value of those awesomeness levels.

- Similarly: 'how awesome would it be if I bring white wine to the party?' Let $H :=$ the proposition that I bring white wine to the party. The answer to the question is this:

$$\sum_W V(w) \cdot i_c(w/H).$$

There we took every maximally specific way things might go, assigned it the appropriate degree of awesomeness, and then took a credence-conditional-on-H-weighted average value of those awesomeness levels.

20.6 Optimization

Should I bring red or white wine to the party? Well, let's ask ourselves how awesome each one would be, and then do the awesomer one: bring white just if

$$\sum_W V(w) \cdot i_c(w/R) < \sum_W V(w) \cdot i_c(w/H);$$

Ties go to red.

What should I do? Well what could I do? Let my options be given in the set O of propositions that I do this, that, the other—for everything I could do. Then: let

$$\max_{A \in O} : \sum_W V(w) \cdot i_c(w/A)$$

be the set of $A \in O$ unexceeded in expected awesomeness by any member of O: for any proposition in that set, I am free to choose it!

21 Classical optimization theory and action primitivism

How to interlace this sort of approach to rationality in action with the theory of knowhow (from the previous handout)? In outline the answer is this: if you know

how to G when F in a lot of different ways, optimize your realization! Formally this gets a little hairy, as you might imagine.

That function stuff was pretty confusing: we can do better.

Trees

Let's think in terms of trees: formally, in terms of the *bounded discrete upper semilattice*.

- A *partial ordering* is a relation on a set of elements which is reflexive, transitive and antisymmetric: $aRb \wedge bRa \equiv a = b$; $aRb \wedge bRc \supset aRc$. We will use $\leq$ conventionally to label partial orderings.
- A *lattice* is a partial ordering for which any two elements have a *supremum* and an *infimum*, where $c = \sup(a,b) := a \leq c \wedge b \leq c \wedge \forall d : (a \leq d \wedge b \leq c \supset c \leq d)$; $\inf(a,b)$ is defined similarly.
- An *upper semilattice* is a partial ordering for which any two elements have a supremum.
- An upper semilattice is *bounded* just if some (unique) element is unexceeded.
- A partial ordering $\leq$ is *discrete* just if, whenever $a < b$: $\exists c : a < c \leq b$ such that $\neg\exists d : a < d < c$; and also $\exists c' : a \leq c' < b$ such that $\neg\exists d' : c' < d' < b$.
- We can say that the elements in a tree are *nodes*. If $a > b \wedge \neg\exists c : a > c > b$, we call a the *mother* of b; b the *daughter* of a. We say also that the maximal node is at *level 0*; and that if a's mother is at level $n-1$, a is at *level n*.

Games

Let's say that a *game* is a tree the nodes of which are *contexts*, subject to the following conditions:

- We assume a set $\mathscr{A}$ of *basic action kinds*. (We haven't gone over to the dark side here: these are just action kinds we are *treating as basic* for the formalism. The story applies again at each lower level, so these actions are not absolutely basic.)
- Suppose that node $c = \langle i_c, \tau_c, \pi_c \rangle$. Suppose that c updates by adding some free choice from $\mathscr{A}$—M, say—to π_c. There are ever so many ways this might play out by the culmination of the M-ing started just then. Some results will be uniform across such new contexts $c*$:

- For every $C \in \tau_{c*}, I^C$ will be simultaneous with the completion of an M-ing;
- π_{c*} will not include M;
- If $w \in |i_c|$, $w \in |i_{c*}|$ only if $I^{\tau_c w}$ is simultaneous with the commencement of an M-ing.

However, we may suppose that, for any subset S of $|i_c|$ subject to these constraints, we may suppose that for some $c*$, $|i_{c*}| = S$. Say that the set of all such contexts is the *M-closure* of c.

[***this could use a little work]

- Then we can say that a tree is a game just if at each node c, its daughters are just the M-closures of c, for every $M \in \mathscr{A}$.

Strategies

Let's say that a *strategy* is a subtree of a game subject to the following restriction: at each node c, its daughters are just the M-closures of c, for *exactly one* $M \in \mathscr{A}$.

Strategies for *G*-ing

Let's say that a strategy $\mathscr{S}$ is a *strategy for G-ing* just if: at each node c, $\langle G, |i_1|, \mathscr{S} \rangle \in \pi_c$; at the top node 1, for all $w \in |i_1|$, $I^{\tau_1 w}$ is simultaneous with the commencement of a G-ing; and at every node c, c is terminal just if for all $w \in |i_c|$, $I^{\tau_c w}$ is simultaneous with the culmination of an G-ing.

Probability of reaching a node

For the nodes in a strategy:

- $\Pi_a(b) = 0$ if $b \not\leq a$; otherwise:
- $\Pi_a(a) = 1$;
- Where a is the mother of b, $|i_a| \setminus |i_b| = P$, and $i_a(P) = x$: $\Pi_c(b) = \Pi_c(a) \cdot x$.

Value as a *G*-ing

We can imagine a value function $V_G(c)$ mapping contexts at which G-ings have just culminated into real numbers: a cardinal measure of the excellence 'qua G-ing' of the just culminated G-ing.

Expected value of a strategy for *G*-ing

We can define the expected value of a strategy for *G*-ing with terminal nodes T as follows:

$$\sum_{t \in T} \Pi_1(t) \cdot V_G(t);$$

the initial credence-weighted average of each way of carrying the strategy forth.

Strategies for *G*-ing under certain circumstances

A strategy for *G*-ing is a strategy for *G*-ing relative to $\langle i, \tau \rangle$ just if for some $c = \langle i_c, \tau_c, \pi_c \rangle$, c is the top node in that strategy.

If a strategy is a strategy for *G*-ing relative to $\langle i, \tau \rangle$, it is a strategy for *G*-ing relative to $\langle i', \tau' \rangle$ whenever $i' \subseteq i$ and $\tau' = \tau \cap i' \times i'$.

Trying to *G*

If one is trying to *G*, one is carrying out a strategy for *G*-ing relative to certain circumstances which might be met. In this case I propose that the initial expected value of such a strategy is *the expected value of succeeding with the strategy* multiplied by *the probability of being in circumstances for succeeding with the strategy*: EV = SV · ES.

Suppose that $\mathscr{S}$ has a higher SV than $\mathscr{T}$: if one were confident of success no matter which one chose, one would pick the former. But if ES($\mathscr{S}$) is low while ES($\mathscr{T}$) is high, the latter might have a higher EV than the former. Diffidence breeds mediocrity.

Part VI
Carnap

22 The semantics-epistemology interface

Let's distinguish the sentence (or string) S from its semantic value (if any) $\|S\|$: the latter a proposition or set of possible worlds, we will assume.

Carnap says:

1. If S is 'testable', S is 'meaningful' (has a semantic value; $\|S\|$ exists);
2. If S is meaningful, S is testable;
3. If any 'test' which 'establishes' S establishes S', the content of S is not exceeded by the content of S';
4. If the content of S is not exceeded by the content of S', any test which establishes S establishes S'.

Why accept these? Well it depends on what we mean by 'testable', 'test', and 'establish', obviously.

Let's say that to establish whether S is to reasonably adopt an attitude of certainty either toward S or toward its negation, $\neg S$. And let's say that S is testable just if whether S 'can' be established: just if there are possible circumstances under which one reasonably adopts an attitude of certainty toward either S or $\neg S$.

What is it to adopt an attitude of certainty toward S? Well suppose that S is a sentence of one's own language, where, if it means anything to one, S means that P. Under these circumstances, for one to adopt an attitude of certainty toward S is for one to update by affirming that P: to take the hypothesis that P into the picture of the world encoded in one's beliefs, to newly restrict the set of worlds one is willing to take seriously to those in which P.

Now, according to our view, one reasonably takes the hypothesis that P into the picture of the world encoded in one's beliefs just when to fail to update by doing so would result in a total picture of the world that is incoherent: for instance, if the picture of the world encoded in one's perceptions and actions, together with the picture of the world encoded in one's beliefs, entails that P.

So, unpacking a bit, the claims read as follows. Suppose that S and S' are meaningful to one and mean, respectively, that P and that P' to one. Then:

1. If there are possible circumstances under which one's total picture of the world entails that P, S is meaningful;

2. If S is meaningful, there are possible circumstances under which one's total picture of the world entails that P;

3. If whenever one's total picture of the world entails that P it entails that P', the content of S is not exceeded by the content of S';

4. If the content of S is not exceeded by the content of S', whenever one's total picture of the world entails that P it entails that P'.

Comments:

1. Now of course we are already assuming that S is meaningful to one, so (1) is a triviality. If we discharge the assumption, (1) requires a connection between S and P if it is to be substantive; obviously only the existing assumption would work. But we might regard a version of (1) in which 'entails that P' is replaced by 'must come to be partly encoded in S' as a bit more substantive; and this seems like a framework assumption of the semantic conception of rationality, on which the rationality of an update is only ever due to its necessity in maintaining coherence.

2. This is more substantive. Perhaps we could justify it on somewhat 'pragmatic' grounds: we should set metaphysics up so that we don't waste time worrying about the pointless.

   But why suppose that there aren't some possibilities among which we just couldn't ever discriminate? An argument might run: we are a lot smarter than cats, and can make discriminations they can't; so why not suppose there could be beings smarter than us who can make discriminations we can't? Or: surely there are either an odd or an even number of galaxies (setting vagueness aside); but we will surely never form an opinion one way or the other.

   Replies: on the first, this shows the view may require a sort of 'pragmaticized' transcendental idealism. On the second, we could read the 'possible' to set the net as coarsely or as finely as we might desire, for whatever purposes we wish to save time.

3. This is a bit like (1): we could regard it as a stipulation of what is meant by 'content'.

4. This is a somewhat natural extension of (2).

23 Psychological categories

Carnap's discussion is hampered somewhat by a vulgar understanding of the psychological categories.

Consider a case in which one forms a belief about the external world:

1. Sam *sees* an indefinite range of objects, parts of objects, and occurrences in which they participate: for example, a certain widget and its states of having a distinctive shape and shade and the event of its being moved up the conveyor belt

2. Sam *turns visual attention on* some selection from among these: for example, the widget's state of having a distinctive shade

3. Sam *forms a visual belief de hoc* with the widget as the standard of truth on the basis of this episode of visual attention: for example, ascribing the property *being defective*

4. Sam *forms a belief de dicto* on the basis of this belief de hoc: for example, affirming the set of worlds in which that very widget is defective

Consider a case in which one forms a belief about one's own body:

1. Brett's body is *qualified by* an indefinite range of sensational occurrences in which its various parts participate: the itch in his toe, the pain in his neck, the sense of cold in his hands and warmth in his feet; the throbbing in his tooth

2. Brett *turns bodily attention on* some selection from among these: for example, the toe's state of having an itch of a distinctive sort

3. Brett *forms a tactual belief de hoc* with the toe as the standard of truth on the basis of this episode of bodily attention: for example, ascribing the property *being dry*

4. Brett *forms a belief de dicto* on the basis of this belief de hoc: for example, affirming the set of worlds in which that very toe is dry

Perceptual copular sentences are true of our subjects in these situations: 'the widget looks defective to Sam' is true—the widget looks defective to Sam—because the widget is the standard of truth for her visual belief de hoc, and its content entails *being defective*; 'his toe feels dry to Brett' is true—his toe feels dry to Brett—because the toe is the standard of truth for his tactual belief de hoc, and its content entails *being dry*.

Belief attributions are also true of our subjects: 'Sam believes that that widget is defective' is true because Sam's belief content entails that that widget is defective; 'Brett believes that his toe is dry' is true because Brett's belief content entails that his toe is dry.

This story is not fully complete: the role of the subject remains obscure, as we will see.

24 Protocol and theory

Protocol sentences function a bit like our 'evidence'. A protocol is an 'epistemic point of departure'; indefeasible; that in which the 'immediately given' is expressed; reasonably treated as a certainty.

I propose to treat the act of attention as the protocol; the rest as theory.

The protocol transfers justification to the belief de hoc by entailing it. The belief de hoc transfers justification to the belief de dicto by entailing it. The protocol does not admit of justification; moreover, it is tacitly known to be true essentially, and is therefore always reasonably treated as a certainty.

24.1 Protocols

'Tacitly known to be true essentially?' For all attendable properties F,

- **Speech act theory:** there is a type of sentence L^F such that to utter it is to turn visual attention to to something's F-ness;
- **Theory of content:** L^F has the de hoc content *F-ness*;
- **Metasemantics:** it has this content because every token of this sentence has the attended state of F-ness as a part (and the general metasemantic rule for the language is 'L^φ has F-ness as its content just if every token of L^φ has F-ness as a part');
- **Theory of truth:** the standard of truth for an utterance of such a sentence is the object of attention.

Suppose that one turns visual attention to o's state of Fness. By the speech act theory, in so doing, one utters L^F. By the theory of content, one is therein affirming the de hoc content *F-ness*. By the metasemantics, this is not due to assignment by context, but is rather an essential feature of the sentence type. By the theory of truth, the utterance is true, because the object of attention is in a state of F-ness (namely the one to which one directs visual attention), and for the utterance to be

true is for the object of attention to be within the content of the utterance—to be F. These are all linguistic rules, which one tacitly knows in the course of turning attention.

Accordingly, L^F has a status something like 'enserfed analyticity'. An analytic sentence is guaranteed by linguistic rules to be true whenever uttered. Standard 'free' analytic sentences (such as sentences of logic and math) are, moreover, freely utterable, and therefore necessarily true. By contrast, our enserfed analyticities are true contingently and temporarily. They get to be true whenever uttered because they are not freely utterable.

24.2 Support of de hoc and de dicto belief

Sentences underlying de hoc belief are distinct from our enserfed analyticities. (Alternatively, we may say that the visual demonstrative 'thus', based on visual attention to F-ness, is the occurrence in de hoc belief of V^F.) After all:

- De hoc belief can be false, protocols cannot

If they are to be linked, such that the latter support the former, we need some translation or inference rules. These would be rules of form

$$\Delta \vdash \varphi,$$

where Δ is a set of sentences including at least one protocol and φ is a sentence of de hoc belief, such that in any context, the content of φ is true just if the content of each member of Δ is true.

Similarly, de hoc belief should support de dicto belief by entailment. For example, the visual de hoc sentence 'Red' should entail the de dicto sentence 'that$_v$ is red', in the sense that in any context in which 'Red' is true, so is 'that is red'. The former is a bit harder so I will put it off.

On the latter (let's ignore time). First on the de hoc side. The context-invariant content of 'Red' is the property *redness*; 'Red' has a truth-value relative to c just if in c, some object is a target of visual attention; if c is such a context and in c, the target of visual attention is o, 'Red' is true relative to c just if o, as it is at the world of c, is in $\|\text{Red}\|$; just if o, at the world of c, is red.

Now on the de dicto side. We may say that the sentence 'that$_v$ is red' expresses a proposition relative to c just if in c, some object is a target of visual attention; if c is such a context and in c, the target of visual attention is o, $\|\text{that}_v \text{ is red}\|$ is the set of worlds at which o is red. Similarly, relative to c, a de dicto sentence is true just if the proposition it expresses relative to c contains the world of c. So, relative to c, 'that$_v$ is red' is true just if o, at the world of c, is red. —Hence, entailment.

Note that on this treatment, both visual de hoc and de dicto visual demonstrative sentences presuppose a target of visual attention: the former in a weaker sense in that such a target is required for a *truth-value*; the latter in a stronger sense: otherwise there is no *content*.

In typical cases (maybe all: see below), visual attention selects from among the seen. Accordingly, the relevant sentences presuppose that one sees; and the de dicto visual sentences presuppose that the subject matter is a seen object.

('No content? That's absurd; butterfly guy (Evans) is still making all sorts of inferences in the normal old way.' Well, note first that these inferences may be able to be backed up at the level of de hoc belief. Note second that 'that$_v$ is red' presupposes 'something I see is red'; which has content. Accordingly in many cases it may be possible to push the rationality of the inferences off to the rationality of the inferences among the presuppositions.)

Inference rules

What do the inference rules from protocol to de hoc belief look like? This demands some subtlety because the content of protocols is contested. Here are some models:

- **A standard inference rule:** $\varphi \supset \psi; \varphi \vdash \psi$
- **Acceptance of an identity sentence:** $\Phi(\text{Hesperus}) \dashv\vdash \Phi(\text{Phosphorus})$
- **An internalist commitment to accept an identity sentence:** the first heavenly body visible in the evening traces a continuous path through the night sky culminating in the same position occupied by the last heavenly body visible in the morning, and $\Phi(\text{Hesperus}) \vdash \Phi(\text{Phosphorus})$; also

  the first heavenly body visible in the evening traces a continuous path through the night sky culminating in the same position occupied by the last heavenly body visible in the morning, and $\Phi(\text{Phosphorus}) \vdash \Phi(\text{Hesperus})$;
- **An externalist commitment to accept an identity sentence:** if the first heavenly body visible in the evening traces a continuous path through the night sky culminating in the same position occupied by the last heavenly body visible in the morning, then: $\Phi(\text{Hesperus}) \dashv\vdash \Phi(\text{Phosphorus})$;

The difference between these last two is that the former binds me only if I accept the claim about the heavenly bodies. In practice, I will *follow* the latter only if I accept it; but I might be falling short of my own standards in failing to do so. (Compare the more general issue of 'wide scope' rationality.)

Hallucination

Let Sam's protocol be V^R. We can characterize the following inference rules:

- **An internalist rule in perception:** blah blah blah and $V^R \vdash$ 'Red';
- **An externalist rule in perception:** if blah blah blah: $V^R \vdash$ 'Red'.

Considerations from hallucination (error more generally) bear on these.

Sam sees the red widget; Hal hallucinates a red widget. They nevertheless accept the same de hoc belief sentence, 'Red'. If Hal came later to learn that he had hallucinated, he would abandon 'Red'; equivalently, Sophie, hallucinating as does Hal but knowing what he might later come to learn, does not accept 'Red'. Conversely, Carla, seeing as does Sam but with Sophie's general view, also does not accept 'Red'.

Same enserfed analyticity throughout? Or different?

Direct realists say it is different. Carnap seems to want to say it is the same (note that this would prohibit saying that the content of Sam's enserfed analyticity is the property *redness*). The former *asymmetric* view is compatible with *thick* content to protocol sentences; the latter *symmetric* view requires them to have *thin* content. After all, protocols must be true, so the content of Sam's protocol would have to be a highly indeterminate property, found both in her situation and in Hal's: presumably something brainy.

Now, if Sam's protocol V^R has thin content, it does not entail 'Red'. Accordingly, no externalist rule would be reasonable. If 'blah blah blah' is set just right, it together with the thin content might still entail 'Red', so that some internalist rule would then be reasonable.

Let us characterize *externalism* as the view that some externalist view is correct; *internalism* as the denial of externalism. We have just seen that if externalism is true, V^R has thick content, and the asymmetric treatment of protocols is correct.

Conversely, if V^R has thick content, externalism is true. A 'defeasibility argument' for thin protocol content plays a big role in Carnap's paper. This argument presupposes internalism; but to grok it we should talk more about what externalists are committed to.

Comments on the internalism–externalism stuff

- We may think that when one comes to see a certain sort of hunk of metal 'as' a carburetor, this involves the acceptance of a new inference rule with the visual attention sentence about that sort of metal in the premiss position and the de hoc belief sentence about carburetorhood in the conclusion position.

If so, then seeing a hunk as a carburetor requires *seeing* carburetorhood, and the capacity to *turn visual attention* to this property. Otherwise there would not be entailment by protocol alone, but only by that together with belief.

Development of seeing-as involves an increase in one's attentional capacities, but not necessarily an increase in determinacy in what one sees: after all, attention is *selective*; perhaps what is going on is that one is selecting new aspects from among what one already saw.

- Internalism may pose rather significant restraints on perception and on thought. If only brainy properties ever show up in protocols, then only brainy properties and brains are targets of visual attention. Plausibly, attentive selection is (relative to capacities and the like) 'free' in regard to what is perceived. If so, only brains are perceived. Preserving our semantic theory for de dicto visual sentences would require accepting that no such sentence ever has the external world as its subject-matter; any sentence 'that$_v$ is a chair' is false (because no one's brain is a chair).

- Suppose that Hal advances a visual de hoc, or de dicto visual demonstrative, judgement. As we have seen, such a judgement presupposes that Hal sees.

- Sophie has a protocol. According to the externalist, it is distinct from V^R. So what is it? Let us suppose that Sophie's fancy knowledge is that she is dreaming in a certain way: a very special way, *red-dreaming*, which she presupposes to be incompatible with the sort of activity involved in seeing a red thing. And let us suppose that in the presence of this knowledge, Sophie 'sees her self as' dreaming in this way: this very special brain activity is sort of 'manifest' to her. If so, this special activity is a target of her visual attention; so the semantic value (and constitution) of her protocol is *red-dreaming*. And relatedly, among her de hoc belief sentences is 'red-dreaming', a sentence ascribing *red-dreaming* to the target of attention: Sophie's brain.

- Carla also has a protocol. Suppose that Carla shares Sophie's general view, and in de hoc belief also ascribes 'red-dreaming' to the target of attention, and presupposes the incompatibility of *red-dreaming* with normal seeing.

  The problem is that whatever her protocol may be, it is something selected from among her situation: some aspect of normal seeing. But now there is an incoherence in her view: her de hoc ascription of *red-dreaming*, her genuine protocol ascribing some aspect of normal seeing, and her presupposition of incompatibility of any such aspect with *red-dreaming*, can't all be true.

  A milder version of Carla faces a milder difficulty: rather than positively ascribing a weird brain property, she simply withholds judgement. Although

she 'as if' directs attention to the redness of the widget she sees, thereby introducing V^R into her protocol, she fails to ascribe 'Red' in de hoc belief.

- Does Hal have a protocol? Hallucinating in the weird manner of Sophie and Carla, Hal is clueless. He directs visual attention as to a red widget he sees, so as to select some aspect of his perceptions as a protocol. In de hoc belief, he ascribes 'Red', presupposing that he sees and that the target of his attention is red. Either he has a protocol or he doesn't. If he doesn't, then his de hoc belief is groundless. But if he does, his presupposition is incompatible with his protocol.

- According to the externalist, there are plenty of ways in which responding to perception can render one irrational, unbeknownst to one: one can wind up in an incoherent condition; or one can fail to form a belief that is rationally required; or one can form a belief groundlessly.

  The reason for this is straightforward. Faced with the externalist rules

  $$\text{If blah blah blah: } \varphi \vdash \psi \qquad \text{if yada yada: if things are } \varphi\text{-ish} \vdash \rho,$$

  the best we can do is follow our beliefs: having accepted φ or at least that things are φ-ish, if I believe 'blah blah blah', I will conclude ψ; but if I believe 'yada yada', I will conclude ρ. My best might not be good enough. Suppose I believe 'yada yada'; I wind up accepting ρ. Sadly, that belief is false: the truth, instead, is 'blah blah blah'. I am, in fact, rationally required to accept ψ.

  Now this tends to freak people out. 'B-b-b-but he's not going around blithering!' Well of course not; the claim is solely that merely bringing rational psychology to the table won't suffice in psychological explanation. It better not be: no one is an internalist about absolutely everything, not even Lewis, not even Chalmers.

  Alternatively, one might be dismayed that one might wind up irrational in response to perception no matter what one does: 'refuse to form the belief and you're not doing enough; form the belief and you're doing too much!' Well no: exactly one of these is the threat, not both. 'B-b-b-but ya can't tell which one it is!' Well sure you can: if you're right, then you can 'tell'.

  Alternatively, one might be dismayed that one can wind up irrational through no fault of one's own. 'The hallucination just came outta nowhere!!' Well no, one only winds up irrational through having false presuppositions: perception is never false, so I don't see who else could be responsible for those false presuppositions!

- Ultimately, if you fancy this sort of indefeasibilism about evidence, you're caught between internalism with its crazy views on perception and externalism with its accusations of irrationality. I'm inclined to think the defeasibilist project isn't going anywhere fast, and indefeasibilism is adequate phenomenologically. Accordingly I'm going with externalism.

24.3 Carnap and internalism

Recall Carnap's 'defeasibility argument' I mentioned above as providing the converse implication from thick content to externalism. This argument shows up in a bunch of contexts in the paper, notably the following. Carnap exploits internalism in arguing against the perception of 'consciousness'—his word for mental attributes thought of dualistically—of the other.

> If we could perceive the sadness of the other it could enter our protocol. Now, any time we form a belief in sadness, we might later discover that the guy was just faking; in which case that belief would get kicked out of our theory. But no protocol can get kicked out. Accordingly, the protocol has thinner content than the belief.

The problem with this argument is that the datum is not that we *discover* the fakery, but that we *think* we did. We might be *wrong* to kick the belief out, and in so doing wind up irrational. For the internalist, whose rules are absolute and apply transparently, this is not an option; for the externalist, whose rules may apply unbeknownst, no problem.

25 Self and other

Although this argument does not work, there are nonetheless a number of self-other asymmetries worth noting:

- It is very plausible that I do not have protocols of the sadness, or pain, of the other. After all, I am simply not wired up so as to be able to turn tactile attention to features of any body other than this one.

  To the extent that pain can only be the subject-matter of a protocol when it is the content of an enserfed analyticity tokened in a focus of tactile attention, the only pain protocols I will ever have are those for this body.

- Sam and Mo, looking at the same state of redness, utter token protocols that are numerically distinct (if kind-identical). Sam's *utterance* of her token

protocol π^* is a token state σ^* that Sam is in, a token state distinct from the one Mo is in.

How isolated is Sam's system of protocols? Sam's state σ^* is, perhaps, bound to Sam: no one else could possibly be in σ^*. Sam's protocol π^* is, perhaps, bound to Sam: no one else could possibly utter π^*. A weaker claim is even more plausible. Sam perceives the world alone: always, if Sam is in a certain token state of uttering a protocol, no one else is in that token state. A still weaker claim is surely true: Sam typically perceives the world alone: ordinarily, if Sam is in a certain token state of uttering a protocol, no one else is in that token state.

So in such an ordinary case, even in the event that Bill does manage to render a judgement concerning Sam's utterance state σ^*, the position Bill occupies in relation to σ^* is very different from the one Sam occupies: Sam is *in* σ^*, Bill is not.

(I am somewhat inclined to think that for one to be in a state of this sort is for that state to be part of one's life, where one and one's life are something like categorial refractions of one another.)

- This asymmetry manifests in an asymmetry in the ways with which Sam and Bill 'appreciate' σ^*: in the ways in which σ^* interacts with each of their pictures of the world. Principle:

  - One is in a state in which token protocol π is uttered just if π 'partly encodes' one's picture of the world;

  Where the rough idea here is that one 'accepts' π, so that it is among the tokens the content of which collectively makes up one's picture of the world; it does so 'transparently'; one is bound to the inference rules governing sentences of its kind; and so on.

  Accordingly, Sam's appreciation of σ^* consists in her affirmation of π^*'s content (and so forth), while Bill's appreciation of σ^* could at best consist of affirming propositions about σ^*. Of course, Bill might affirm the content which happens to be the content of π^*, while Sam might affirm propositions about σ^*. But the former would not be sufficient for Bill to appreciate σ^* in the same manner as does Sam, and the latter is not necessary for Sam to appreciate σ^* in the manner in which she does in fact appreciate it.

  We might express the difference by saying that σ^* *underlies* Sam's point of view but is *within* Bill's point of view.

We might also say that different *aspects* of σ^* show up to Sam and to Bill. We could name these aspects: the one showing up to Bill is σ^*'s *external* aspect; to Sam, its *internal* aspect.

Or we could characterize this difference in terms of *perspectives*: Bill takes a perspective *on* σ^*; Sam takes a perspective *from* σ^*.

- Bill cannot literally access σ^*'s internal aspect, adopt a perspective from σ^*. However, if Bill is in a state of the same kind as σ^*—if, like Mo, Bill too tokens a protocol of π^*'s kind—his picture of the world will share a certain similarity with Sam's: both will affirm the de hoc content *redness* in a way bound to a certain class of inference rules.

  More expansively, if two people tokened exactly the same kinds of sentences, their pictures of the world would be the same: they would picture the world in the same way; the way the world is in one picture is the same as the way it is in the other.

- Similarly, via 'simulation'—perhaps, imaginative exercise—Bill can create 'within himself' a state of a similar kind to σ^*. This activity would result in his adoption of a sort of 'picture-in-picture of the world' which pictures the world in a manner similarly to the way in which Sam pictures the world.

  This is what Carnap is talking about in his discussion of 'Verstehen' and 'meaningful behavior'. Suppose that Rich is in pain: he is in a state σ^{**} of uttering a protocol about *painfulness* and thereby ascribing pain in bodily attention. Rich behaves accordingly. Jane sees Rich's behavior. But she doesn't just see the behavior: she sees it *as* pain-behavior.

  On our view, what this amounts to is Jane's simulating Rich's pain: Jane's creating—or finding—within herself a state similar to σ^{**} and thereby adopting a picture-in-picture of the world ascribing pain in bodily attention.

  How does this state arise? The literal content of Jane's belief can involve no more than an ascription to Rich of uttering a protocol about painfulness—from a perspective *on* this property. Perhaps we are 'wired' in such a way that any belief with this content causes one to adopt a picture-in-picture that ascribes pain in bodily attention.

  Alternatively, perhaps Jane's belief does not cause Jane's picture-in-picture: perhaps the relation is rather that the vehicle of Jane's belief and the vehicle of Jane's picture-in-picture are one and the same.

  We will talk later about the place of this issue in discussions of physicalism.

- Self-other asymmetries aren't limited to those found at the level of enserfed de hoc belief sentences.

  We are assuming that free de hoc beliefs and de dicto beliefs also have their contents encoded in 'sentence-like' entities (I'm not thinking in terms of any language of thought or coggish or boxological stuff: merely that there is some metasemantical basis for belief content, perhaps behavioral yada yada).

  This assumption allows us to generalize the discussion above. When I am in a state of uttering a certain sentence in de dicto belief, I appreciate the internal aspect of this state, take the perspective from it. Anyone else can only appreciate its external aspect, take a perspective on it.

  Accordingly, we can simulate beliefs of all stripes: create within ourselves states of the right sort and we add the content of the simulated belief to our picture-in-picture.

  Then when we see someone as acting out of certain beliefs, that involves our being in a substate of the kind with content similar to that of those beliefs. That substate is caused by a belief ascribing to them the utterance of a sentence of the sort that is a vehicle for that content; or perhaps the vehicle of that belief and the vehicle of the substate are one and the same.

- This story is not limited to the theoretical aspects of the mind.

  If Sam is *G*-ing, there is a certain token process γ, a *G*-ing, of which Sam is the agent. (In my view, for Sam to be the agent of γ is for γ to be part of Sam's life, where Sam's life is sort of the same thing as Sam.)

  For Sam, the internal aspect of γ shows up; for others, at best its external aspect can show up. Sam takes the perspective from within γ; others at best a perspective on γ.

  Someone can simulate Sam's *G*-ing by setting up a fake mini process of a similar kind to γ within themself. In that case, then they can take something like the perspective from within *G*-ing.

  When Bill sees Sam as *G*-ing, that involves Bill's being in a substate which is a *G*-ing; perhaps this substate is caused by a belief ascribing *G*-ing from without; perhaps the vehicle of the substate and that of the belief are one and the same.

- Let us draw some lines of comparison and contrast between the theoretical and practical aspects of the mind. I begin by making explicit the analogy.

The vehicle of a theoretical state—a belief—is a state of uttering a certain kind of sentence; the kind of vehicle is given by the kind of sentence. That is what shows up from without. What shows up from within is affirmation of the content of the kind of sentence, inclusion of that proposition within one's picture of the world. This is the familiar 'use-mention' distinction.

What shows up from without in the case of an action is a process of a certain kind, perhaps a behavioral kind. What shows up from within is the action-kind itself, as performed. This is what the use-mention distinction looks like in the practical domain.

An initial respect of disanalogy. Note that in the theoretical domain, there is some distance between the subject and the respects of individuation both from without and from within. All belief-vehicles involve *utterance*; they differ in respect of which *sentence* is uttered. All episodes of affirmation involve *affirmation*; they differ in respect of which *content* is affirmed. (Recall Moore's remarks on that which makes the sensation of blue a mental fact.) By contrast, we observe no such distance in the practical domain. This makes phenomenological sense: I am not the world, I merely picture the world; by contrast, I perform my actions.

A second respect of disanalogy. The theoretical use-mention distinction is familiar, the practical use-mention distinction unfamiliar. Why is this? Perhaps because vehicles of belief can be externalized and made objects of collective scrutiny as sentences of public language. You and I observe the same written sentence, attempting to ascertain its content. By contrast, vehicles of action are necessarily private. Though perhaps the doctrine that experience is the best teacher is an implicit recognition of the practical use-mention distinction.

- I claim that the perspective from within is the one encoded in central ordinary language psychological vocabulary: perceptual copular verbs, 'believes', and action predicates.

  Consider our semantic theory for belief avowals: 'I believe that P' tests the context given by the speaker's psychological profile for whether the content of de dicto belief entails $\|P\|$. This theory has two consequences: first, that the the logical form of this sentence does not include an argument place for the subject; and second, that the speaker must accept it just if he or she accepts 'P'. This connection to acceptance of the complement sentence suggests that the picture encoded in 'I believe that P' is in some way equivalent to that encoded in 'P'. This in turn suggests that the function of the belief avowal is to in a sense 'show' one's perspective as being that of acceptance

of '*P*'. That is, of course, the perspective from within. In light of this, the absence of an argument place for a subject is no surprise: I can only (genuinely, as opposed to in simulation—more below) take the perspective from within on a single subject—myself. Inclusion of an argument place would overgenerate, requiring an ability to (genuinely) hop perspectives that, well, doesn't make any sense.

Consider our semantic theory for perceptual copular verbs: 'this$_v$ looks *F*' expresses a proposition just if there is some target of visual attention; if it expresses a proposition, it tests the context for whether the content of visual de hoc belief entails $\|F\|$. Same points here.

Consider our semantic theory for action avowals: 'I am *G*-ing' tests the context for whether $\|G\|$—*G*-ing—is on the list of practical matters—for whether one is in fact *G*-ing. Since one must accept this sentence just if one is *G*-ing, the perspective it encodes is that of the *G*-er: the one who is in fact the behavioral vehicle of a *G*-ing. Similar point for the absence of argument place.

- Evidently then psychological *ascriptions*—'Sam believes that *P*', '*o* looks *F* to Sam', 'Sam is *G*-ing'—would encode the perspective from without. How do they do this? The story, as is very well known, must be sensitive to a range of data:

  - It is important to preserve the sense that when one says 'Bill believes that *P*', one says of Bill what one says of oneself when one says 'I believe that *P*'. In a slogan, *belief ascriptions are displaced belief avowals*. The alternative would be to posit a change of subject between avowals and ascriptions.
  - As a consequence of this, the meaning of a belief ascription must involve the content of the embedded clause. 'Sam believes that *P*' does not represent her as a purely syntactic engine, but rather requires her to have a picture of the world, one entailing something like $\|P\|$.
  - And yet it is important not to make the meaning of an ascriptive belief-predicate vary solely with the content of the embedded clause. We want to allow for distance in meaning even between ascriptions with intensionally equivalent embedded clauses. So it would seem as if the theory would involve some degree of 'mention'.
  - In incorporating mention into the theory, it is important to accommodate extensive departures from literalism: a belief-ascription, famil-

iarly, will often embed a sentence unfamiliar to the subject of the ascription.

– And something similar goes for the use side: the embedded sentence in the mouth of the ascriber will almost always have a different meaning in the mouth of the ascribee.

These data can be accommodated by expanding on a proposal we canvased earlier. Preliminarily, let us say the following:

– The logical form of 'Sam believes that φ' is $\mathsf{FROM}^{s_{\sigma_c(\mathrm{Sam})}}\mathsf{B}\varphi$;
– $\lceil \mathsf{FROM}^{s_j}\varphi \rceil^c =$
  * $\lceil \varphi \rceil^{c/s_j}$;
  * $c/s_j = \langle w_c, n_c^j, \langle r_c, n_c^1, n_c^2, \ldots \rangle \rangle$;
– σ_c matches a subject with the number of the sequence position being used in c to keep track of her information state.

To see how this works, suppose that sequence position 1 is used to keep track of Sam's information state. Then the lf of 'Sam believes that φ' is '$\mathsf{FROM}^{s_1}\mathsf{B}\varphi$'; this in turn expresses at c the proposition $\lceil \mathsf{B}\varphi \rceil^{c/s_1}$; which in turn represents the outcome of the test for whether i_{c/s_1} entails $\lceil \varphi \rceil^{c/s_1}$.

Suppose then that φ is a simple sentence such as 'snow is white'; in that case, the proposition expressed by φ is context-invariant, so that the test is passed at c just if $n_c^1 \subseteq \lceil$Hesperus is a planet$\rceil$; just in case Sam is represented in c as being such that her picture of the world entails that Hesperus is a planet.

Accordingly, if (note: not *only* if) one accepts 'Sam believes that Hesperus is a planet', one represents Sam as being such that her picture of the world entails that Hesperus is a planet; and a cooperative audience will conclude on the basis of an assertion of this sentence also that one does so.

It is clear, I hope, that this theory goes some way toward capturing the idea that when I say of Sam that she believes that *P* I say of her what I say of myself when I say that I believe that *P*: belief ascriptions are displaced belief avowals.

But it still leaves a fair bit unexplained. First, it says nothing about the truth-conditions of such a sentence. Obviously I can represent Sam in any range of ways; but we think that some of those are accurate, others are inaccurate; and we think that belief ascriptions are candidates for accuracy and inaccuracy. There must be more going on than Second, being pitched so far purely at the level of content, it does not address the 'hyperintensionality' of belief

ascription (thus the absence of a biconditional above.) Third, it makes no provision for departure from literalism.

To address these points, we need to return from semantics to psychology. What is it that grounds a certain sequence position's being used in a context to keep track of Sam's information state? Suppose that one is running a number of simulations at once: one of Sam, one of Mo, one of Brett, and so forth (maybe we can't do more than one 'in the foreground' simultaneously but we can surely do several if some are allowed to run 'in the background'). In that case, one has a number of substates: one for Sam, one for Mo, one for Brett, and so on. That in the psychology which is mirrored in the semantics by the range of sequence positions is this sequence of substates; the substate for Sam, we are assuming, is assigned to the first sequence position in the semantics.

Now, familiarly, one might represent Sam as having a picture of the world entailing that Hesperus is a planet in at least two ways: the substate for Sam may utter 'Hesperus is a planet', or it may utter 'Phosphorus is a planet'. My suggestion is that the sense in which ascriptions are displaced avowals runs deeper than the semantic: it is not just that the semantic theory for the corresponding avowal is called upon in evaluating the ascription. It runs still further down, to the syntactic level: my utterance of the ascription is an utterance 'as Sam' of the corresponding avowal. The embedded clause in the belief-ascription, therefore, is the sentence uttered in my substate for Sam.

Next, recall that the state in which I represent the view from without on Sam's utterance state is my state of being in a substate for Sam. In being in a substate for Sam which utters 'Hesperus is a planet', I am thereby in a state of uttering a sentence, the content of which is the view from without on Sam's state of uttering something like 'Hesperus is a planet'; if my substate for Sam rather utters 'Phosphorus is a planet', the content of my main state is the view from without on Sam's state of uttering something like 'Phosphorus is a planet'.

How shall our semantic project avail itself of these psychological aspects of belief ascription? One possibility would be to appeal to a semantic 'quotational' aspect of the theory: somehow 'Sam believes that Hesperus is a planet' performs a test not just on the content of my Sam substate but also on its form. How to implement this idea? FROM functions solely to manipulate the information coordinates of the context; and while B has more semantic punch, it would be unattractive to insist that it performs a test not just on content but on form (this is not motivated in the first-person case,

and is indeed unattractive: the performance of such a test would be a redundant excrescence, and would undermine the transparency of consciousness; to maintain parity, all action predicates would need to be loaded up with syntactic tests).

Better instead to appeal to something more like presupposition. When I assert 'Hesperus is a planet', I represent not only that the world is such that Hesperus is a planet, but also that I accept the sentence 'Hesperus is a planet'. In that sense my public utterance presupposes this fact about myself: without that fact, my utterance is at best a lie and at worst an impossibility. When I *accept* 'Hesperus is a planet', I also represent that the world is such that Hesperus is a planet; and while I do not conversationally implicate that I accept 'Hesperus is a planet', the presupposition is deeper: it is metaphysical, in the sense that in accepting 'Hesperus is a planet' I am, well, accepting 'Hesperus is a planet'.

Analogously, when I assert 'Sam believes that Hesperus is a planet', I represent that I accept in Sam's voice 'I believe that Hesperus is a planet': namely that in my Sam substate, I utter 'I believe that Hesperus is a planet'. And when I accept 'Sam believes that Hesperus is a planet', I do in fact accept in Sam's voice 'I believe that Hesperus is a planet'. The public assertion conversationally implicates my acceptance in Sam's voice; the private acceptance is constituted by that acceptance. In this sense, either way, I presuppose that my Sam substate utters 'I believe that Hesperus is a planet'.

But since when my Sam substate utters 'I believe that Hesperus is a planet', I thereby represent whatever content goes with the view from without on Sam's state of uttering something like 'I believe that Hesperus is a planet', I also presuppose this content. And since uttering this renders it rationally compulsory to utter 'Hesperus is a planet', I also presuppose and therefore represent whatever content goes with the view from without on Sam's state of uttering something like 'Hesperus is a planet'.

Let us introduce some notation for talking about this sort of presupposition. The sentential (or discourse) operator $\mathsf{TO}^{(\cdot)}$ is 'quotational', in the sense that the meaning it returns when applied to φ is a function of φ rather than of $\ulcorner\varphi\urcorner$:

- The logical form of the presupposition of 'Sam believes that φ' is $\mathsf{TO}^{\mathrm{Sam}}\varphi$;
- $\ulcorner\mathsf{TO}^{\mathrm{Sam}}\varphi\urcorner^{c}$ is the set of worlds at which Sam is in a state similar (by the standards of c) to the utterance of φ.

Here is a convenient guide to our theory:

– In uttering, in c, 'Sam believes that φ', one affirms the proposition

$$\ulcorner \mathsf{FROM}^{s\sigma_c(\mathrm{Sam})}\, \mathsf{B}\varphi \urcorner^c$$

and presupposes the proposition

$$\ulcorner \mathsf{TO}^{\mathrm{Sam}} \mathsf{B}\varphi \urcorner^c;$$

– The affirmation is the outcome of a test of the context for whether the informational state assigned to Sam entails the proposition expressed by $\mathsf{B}\varphi$ relative to the context and therefore also the proposition expressed by φ relative to the context;

– The presupposition is the set of worlds at which Sam is in a state similar (by the standards of the context) to the utterance of φ.

Set in this way, we can see that the theory resolves our difficulties straightforwardly. The question of literalism is resolved, on the syntactic side, by the 'contextual similarity' aspect of the presupposition; on the semantic side, by the fact that, except via the syntactic side, the semantic side is unconstrained by the world. The questions of truth-conditions and of hyperintensionality that plagued the purely semantic theory are dealt with by the presupposition. Sam is not treated as a purely syntactic object, because the literal content of the sentence involves my seeing things a certain way from her point of view; the metaphysically necessary link of this way to the syntactic side constraints the legitimate ways of seeing things from Sam's point of view.

- My view implies that *there are no propositions of psychology*. The only propositions in the story concern the sentences from within which psychology appears. The psychology in the story runs on solely at the level of the nonpropositional: through my own avowals, and my avowals in the voice of others.

- The truth in solipsism is that in my world, my perspective is unique: it is the only one I can target with avowals without the need to engage in a shift to a perspective which is not mine, which is absent to me. (Analogously, the truth in presentism is that those at the crest of the wave of being can only have access to their past cases of standing in this position by a shift to a perspective which is now gone.)

Nonliteralism does not look like a surprise, once we have taken note of this point. The central aim of psychology is not the physiological aim of characterizing the objective or syntactic side: that is useful only in our attempt to access the inaccessible. But since the psychological is inaccessible, there can be no possibility of accuracy to it.

- Carnap's remarks on 'Verstehen' bear some similarity to the foregoing.

  Theorists of the Verstehen point to the phenomenon of seeing a behavior as an action: in Carnap's example, of seeing a head movement as a nod of affirmation. Carnap acknowledges that one who does so affirms the protocol 'that is a nod of affirmation', which is neither a physical sentence nor one we yet know how to translate into a physical sentence. But in Carnap's view this is no threat to his doctrine of physicalism (about which more below).

  His discussion exploits an analogy to the use of a detector. Suppose that a physicist has a Geiger counter: its ticks are more rapid when the intensity of the ambient radiation is higher. In the presence of a rapidly ticking Geiger counter, her protocol is 'the Geiger counter is ticking rapidly', but she issues the 'system sentence' 'the ambient radiation is intense'. Or suppose that a doctor has a disease sniffing dog: the dog barks in the presence of tuberculosis, remains silent otherwise. When the dog barks, the doctor's protocol is 'the dog is barking', but he issues the system sentence 'the patient has tuberculosis'.

  A more ignorant doctor might issue a 'quantified' system sentence 'whatever that undesirable internal condition going around is that ordinarily causes the dog to bark, the patient has it'. Or he might simply add to his system the sentence 'the dog is barking'—a sentence implicitly treated as equivalent to the quantified sentence. As we will see, in Carnap's view, the legitimacy of none of these practices would establish the denial of physicalism.

  Carnap suggests that when one undergoes an episode of the Verstehen, one 'is both the observer and the detector' (184): one treats oneself as one's own disease sniffing dog. I think he means the following. Suppose that I issue the protocol 'that is a nod of affirmation': here I am in the role of disease sniffing dog, the detector. In doing so, I also issue the 'metaprotocol' 'I issue the protocol 'that is a nod of affirmation' ': here I am in the role of the doctor, the observer (this stage is interpolated: Carnap slurs this point over, though it seems to be implicit in the remark that one 'establishes ... a condition which can be characterized only indirectly—by his own diagnostic reaction': 185). I then add to my system 'whatever that internal condition is that causes me to issue the protocol 'that is a nod of affirmation', that person is in it'.

If this system sentence about my protocols doesn't undermine physicalism (and Carnap thinks it doesn't), then the analogy to detectors more generally shows that there is no threat to physicalism.

The suggestion that in Verstehen one is both observer and detector anticipates our views about the duality in psychological ascription. The issuance of the protocol, one's role 'as detector', analogizes to *being* in the substate; the issuance of the metaprotocol, one's role 'as observer', analogizes to the *semantic* aspect of the psychological ascription; and the issuance of the system sentence analogizes to the *presuppositional* aspect of the psychological ascription.

26 Physicalism

Carnap characterizes physicalism as the doctrine that the physical language is 'universal and intersubjective': *universal* in the sense that any of my true system sentences are translatable into truths in the language of physics; *intersubjective* in the sense that what goes for me goes for everyone. The physical language, in turn, is characterized as having a syntax in which magnitudes are assigned to points of spacetime.

26.1 Physical questions

We, by contrast, think of physicalism as the adoption of a sort of 'stance'. Let us say that one *regards Q as a physical question* just when, of any two distinct cells in the Q-partition of the worlds one regards as coherent, there is a physical difference between any two worlds, one from one cell and one from the other.

Just what it is for two worlds to differ physically is a fine question. Let us consider some examples:

- 'How many electrons in this box?'—that is a physical question, because the cells in the partition are the worlds with 0 electrons in the box, the worlds with 1 electron in the box, the worlds with 2 electrons in the box ... . Any worlds with n versus $m \neq n$ electrons in the box differ physically—in respect of how many electrons are in the box.

- 'Is Fred's arm broken?'—that is very likely a physical question, because the cells in the partition are: the worlds where Fred's arm is broken; the rest of the worlds. Plausibly, a broken arm has a different pattern of electromagnetic bonds than a nonbroken arm: thus a physical difference.

- 'Is there a ghost in the basement?'—that would seem not to be a physical question, because (according to my understanding of what it is to be a ghost, anyway) a ghost is not made of matter and need not interact with matter. The cells in the partition are the worlds in which there is a ghost in the basement and the worlds in which there is not a ghost in the basement. Pick a world from the latter cell; if the concept of a ghost is coherent, we can add a ghost to the (relevant) basement of that world without any alterations in the matter in that world. The resulting world would be in the former cell.

- 'Is this leaf green?'—a harder case. The partition here is the pair of those worlds in which it is, and those worlds in which it isn't. Can we take one of the latter worlds, make a purely nonphysical alteration, and wind up at one of the former worlds? The question is nonphysical just if we can.

  How might we do that? Consider the alleged nonphysical properties 'primitive green' and 'primitive red'. Our strategy would be to change the leaf from green to red by changing it from primitive green to primitive red. Then two jointly sufficient conditions of adequacy:

  1. Are worlds containing primitive green and primitive red coherent?
  2. Would changing a leaf from primitive green to primitive red suffice for changing it from green to red?

  Answer both questions in the affirmative and our strategy would work: the question would be nonphysical. By contrast, to affirm the physicality of the question would be to regard any such strategy as doomed: to regard any nonphysical alteration that could be coherently imagined as insufficient to change this leaf from green to nongreen; to regard the conditions as in each case not both met.

  Suppose this leaf is in fact green. And suppose we encounter a physicalist about whether this leaf is green who takes a sort of ecumenical attitude toward what is coherent. This person's attitude would be that possibilities physically alike but differing in regard to whether this leaf is primitive green are coherent, it is incoherent to suppose that things are physically just as they are but this leaf is not green. For such a person, the physical truth entails that the leaf is green.

  My physicalist here is an 'a priori' physicalist. What to do about the 'necessary a posteriori'? Two strategies:

  – Conceptual change/internalism: before the discovery of the constitution of greenness, we regard some strategy as adequate; afterward we

change our minds;

- Externalism: before the discovery of the constitution of greenness, our assessment of some strategy was 'I don't know'. The reason we did this was that we were uncertain about the status of some wide scope condition that would force our hand one way or other; after the discovery we determined that we were rationally bound to be negative about every strategy.

Each of these strikes me as credible.

- 'Is Sam conscious?'—here the story is much the same. Is there any nonphysical alteration that would suffice for Sam's switching between consciousness and nonconsciousness? This would be so only if, for some feature Sam might be alleged to go from having or lacking without a physical alteration in the world:

  1. Sam's having it and lacking it are both coherent possibilities;
  2. Whether Sam has it or lacks it suffices for altering whether Sam is conscious.

26.2 Physicalism tout court

What is 'physicalism' more generally?

- Perhaps we can group questions together into 'subject-matters': we might think of 'the mental' as a grouping of questions about whether this or that person has pain (believes that *P*, is conscious, is *G*-ing, etc.). Such a grouping would be a question which entails every other question in the group: 'the question of the mental', for example. Then to be a 'physicalist about the mental' would be to regard the question of the mental as a physical question.

- What is physicalism tout court, the global stance?

  One proposal: physicalism is the attitude that every question in a certain valorized class—the important questions, the interesting questions, the significant questions—is a physical question.

  Is this right? Physicalism seems to be affiliated with the rejection of ghosts and the like: the actual taking of positions on certain nonphysical questions. So it seems as if physicalists regard at least some nonphysical questions as interesting.

> The characterization may be nevertheless on the right track. An 'ontological' question like 'are there any ghosts?' plays a special role. A 'characterizing' question like 'are the ghosts happy?'—more generally, a question about *how* the ghosts are—presupposes an affirmative answer to the ontological question. One could therefore get rid of all characterizing questions about ghosts in one fell swoop by rejecting the presupposition: and if one did so, one would regard all those characterizing questions as uninteresting.
>
> So perhaps we could say that physicalism tout court is the stance according to which, whenever the ontological presupposition of a characterizing question is met, it is a physical question.

26.3 Against the dualism debate

That organism over there utters a certain protocol sentence. That widget looks red to Sam. In our view, these are two perspectives on the same reality: the former is more tied to the view from without the utterance, the latter more to the view from within. The latter presupposes the propositional content of the former, while testing the speaker's psychological state for the presence of a substate attached to Sam in which the widget's looking red is simulated.

I characterized a consequence of this view up above as that *there are no propositions about consciousness*. If not, no two coherent possibilities differ with respect to whether that widget looks red to Sam *at all*, let alone despite being the same physically—not if 'that widget looks red to Sam' is understood as having 'intrinsic' propositional content.

Perhaps questions about the utterance of protocol sentences are 'intrinsically' nonphysical: perhaps to utter a protocol sentence involving red requires red, where the question of color is a nonphysical question. But the proposition that a certain organism utters a certain protocol sentence does not concern consciousness. None does.

If this is correct, there can be no question about the status of dualism: not if dualism is understood as the view that some proposition about consciousness is not entailed by any class of propositions about the physical, antidualism as the status that every proposition about consciousness is entailed by some class of propositions about the physical. The debate between dualists and antidualists presupposes that there are propositions about consciousness. If there aren't any, the debate can't take place.

This does not establish that physicalism tout court is a reasonable stance. Our rejection of the dualism debate leaves open whether the question 'which protocol sentence does Sam utter?' is physical. As we have just seen, it may not be.

26.4 The hard problem

The 'problem of consciousness' is the question 'how or why does 'technicolor phenomenology' emerge from 'soggy grey matter'?' This problem is widely thought to be distinctively 'hard'. Is it?

- The hard problem gets started with zombies. One's zombie twin is a creature just like one physically but lacking in consciousness: for whom 'all is dark inside'. Allegedly, one's zombie twin is coherently conceivable: we regard a world containing such a being as a coherent possibility.

  Grant arguendo that zombies are coherently conceivable. If so, this would plausibly show that one's physical nature does not conceptually entail that one is not a zombie. Since one is not a zombie, this fact is not conceptually entailed by the physical facts.

  If explanation requires conceptual entailment, this shows that the fact that one is not a zombie is inexplicable by the physical facts: that there is an 'explanatory gap' between the physical and consciousness.

  If there is an explanatory gap between the physical and everything else, this would not be a big deal: the problem of consciousness would be no harder than the problem of, say, how wet water emerges from elementary physical particles.

  But, allegedly, there is not an explanatory gap between the physical and anything else. Fix the facts about the elementary physical particles and the facts about water cannot be coherently varied. The problem of consciousness is uniquely hard.

  The asymmetry here arises out of the distinctive rules of rationality governing thought about consciousness. For most natural kinds, some externalist wide-scope rule is plausible: e.g.,

  - If the predominant local liquid is H_2O:

    $$\Phi(\text{water}) \dashv\vdash \Phi(H_2O)$$

  With this sort of rule we can explain why the prescientific were not certain that the H_2O facts determine the water facts: they were not, because they were not in possession of the fact that triggers the pattern of equivalences. Accordingly, though they were unaware of being rationally bound to that pattern of inference rules, they were in fact so bound.

  But this externalist rule seems to cede to the world rational control over expression 'water' in a way suggestive of a lack of transparent understanding

of its referent. Accordingly, for entities for which our concepts of which do provide a transparent understanding, an externalist rule would not be plausible. And we are in fact 'acquainted' with consciousness and its various determinates; so our concepts of these properties do provide a transparent understanding of them; so no externalist rule is plausible. The apparent coherence of zombies therefore cannot be explained by anything other than their genuine coherence; this unique genuine coherence makes the problem of consciousness uniquely hard.

- The progression from the coherent conceivability of zombies to the unique hardness of the problem of consciousness is compelling. (Modulo our indecision about wide-scoping versus conceptual plasticity. Modulo also our 'acquaintance' with determinates of consciousness: it is certainly true that such concepts are unlike other concepts, but not in this way. Set this aside, however: if there are propositions about consciousness, the theory of acquaintance provides a fine elucidation of this difference.) If zombies are coherently conceivable, the problem of consciousness is uniquely hard. Are they?

- What would a zombie be on our view? What is the proposition I am asked to conceive of when I am asked to conceive of the existence of a zombie?

  It can't be the proposition that there is a being of whom all the propositions about the physical true of me are true while some of the propositions about consciousness true of me are false: there are no propositions about consciousness.

  What else? If 'all is dark inside' of a certain being, presumably that being utters no protocol sentences. After all, an utterance of a protocol sentence has a perspective from within. Presumably the claim that all is dark inside a certain being is to be understood as requiring that no aspect of the being has a perspective from within. The proposition that there is a zombie would have to be the proposition that there is a being just like one physically despite uttering no protocol sentences.

- So the hard problem, if it is to be motivated by the coherence of zombies, would have to be the question of how protocol sentences can emerge from soggy grey matter. Could this question be uniquely hard?

  Notice that this question is asked from the perspective from without. This perspective does not afford the sort of 'acquaintance' alleged to ground the unique hardness of the problem of consciousness. My difference from my zombie twin is a difference in the perspective from within: I have one, he

doesn't. So this question could not be the hard problem: not, at least, if its unique hardness is correctly diagnosed.

So the proposition that there is a zombie is not the proposition that there is a being physically just like me but uttering no protocol sentences.

- I conclude that the notion of *the proposition that there is a zombie* is incoherent.
- Let us see why that is. This notion is supposed to apply to a proposition ...

  1. Which bears on the existence or character of someone's perspective from within;
  2. And encodes our acquaintance with its subject matter.

  Fail at (1) and the conceivability of zombies could not bear on the problem of *consciousness*; fail at (2) and the conceivability of zombies could not show this problem to be uniquely *hard*.

  The problem is that the only propositions bearing on the existence or character of someone's perspective from within are propositions at the level of *mention* of protocols; but only propositions at the level of *use* of protocols encode acquaintance.

- The appearance that there is a hard problem of consciousness is due to a conflation of use and mention. Its resilience should be no surprise. Use and mention are ferociously difficult to tease apart.
- What do we *do* when we seem to conceive of a zombie? Let's address an easier case, the 'invert' physically just like me but looking at green rather than red.

  My hypothesis is this: we draw up a sentence concerning the doings of some physical being, Inez, viewed from without. This would be a sentence along the lines of $\mathsf{TO}^{\text{Inez}}[\mathsf{B}(\text{that's red})]$.

  We also draw up the negation of the corresponding sentence about Inez's view from within: $\mathsf{FROM}^{\text{Inez}}[\mathsf{B}(\text{that's green})]$.

  We then conjoin these two, yielding:

  $$\mathsf{TO}^{\text{Inez}}[\mathsf{B}(\text{that's red})] \wedge \mathsf{FROM}^{\text{Inez}}[\mathsf{B}(\text{that's green})].$$

  That, if anything, would be the invert sentence.

But (supposing physicalism about protocol sentences) the left conjunct here is incompatible with a presupposition of the right conjunct (under the supposition that you can't have the physical bases of looking at red and looking at green at the same time: enrich the example to get something in the right spirit that is true if you need to). So this sentence does not represent a coherent possibility. The incoherence is a subtle one, involving distinct codings of the same content and a bit of testing. On the left, the coding takes the form of the view from without; on the right, it takes the form of (assuming the test is passed) the pseudo view from within.

This is a Frege case. But it is not exactly a standard Frege case. 'Hesperus is a planet' and 'Phosphorus is a planet', perhaps, differently code the same content. The sameness is obscured due to a wide scope inference rule governing the expressions. Both codings however involve the view from without their subject matter (if views at different times of day).

We may suppose that the sameness in content of $\mathsf{TO}^{\mathrm{Inez}}[\mathsf{B}(\text{that's red})]$ and $\mathsf{FROM}^{\mathrm{Inez}}[\mathsf{B}(\text{that's red})]$ is also obscured by their governance by a wide scope inference rule. It is for this reason that the invert sentence is not transparently incoherent, despite having an incoherent content.

But the views from which their contents are encoded is dramatically different. The former encodes the view from without, but the latter encodes the view from within.

If this is right, then governance by a wide scope inference rule is compatible with the view from within.

- At least we have no reason to suppose otherwise. One might at this point take a modus tollens: I assumed physicalism about protocol sentences. If this requires compatibility between wide scope and the view from within, one might try to pitch that assumption.

  Here David Lewis's judgement is apposite:

  > Let *parapsychology* be the science of all the nonphysical things, properties, causal processes, laws of nature, and so forth that may be required to explain the things we do. Let us suppose that we learn ever so much parapsychology. It will make no difference. Black-and-white Mary may study all the parapsychology as well as all the psychophysics of color vision, but she still won't know what it's like. Lessons on the aura of Vegemite will do no more for us than lessons on its chemical composition. And so it goes. Our intuitive starting point wasn't just that *physics* lessons

couldn't help the inexperienced to know what it's like. It was that *lessons* couldn't help.

If nonphysicalism about protocol sentences won't shut the explanatory gap, pitching the assumption of physicalism about protocol sentences won't restore the entailment from the view from within to narrow scope. We would have to conclude that there can be no objective basis for the view from within, physical or nonphysical. But that's absurd.

The only legitimate rebuttal here is an appeal to the unnameable, to objects of acquaintance, as the subject matter of the left conjunct. But, as we saw above, acquaintance is no part of the view from without.

- The zombie hypothesis looks like this:

$$\mathsf{TO}^{\mathrm{Zeke}}[\mathsf{B}(\text{that's red})] \wedge \forall \Pi : \neg \mathsf{FROM}^{\mathrm{Zeke}}[\Pi];$$

that is to say, Zeke is physically just like someone looking at a red thing but I have no substate for Zeke.

Here the incoherence is more subtle. The invert sentence is like 'Hesperus is a planet but Phosphorus isn't'. The zombie sentence is more like 'Hesperus is a planet, but Phosphorus? Dunno'. In considering an invert, one contradicts oneself, does more than one should; in considering a zombie, one fails to draw out an entailment, does less than one should. Nevertheless, in both cases, one both affirms and fails to affirm one and the same content.

- A slightly more psychological version of the story would run like this. We are generally good at exploiting sympathy when actually looking at someone or reading an action-laden characterization of their doings: then, we take up the view from within swiftly and accurately. We are less good at doing so when the person is characterized in more basic terms: physiologically or even physically. A context concerned with the status of microphysical supervenience puts us at a grave disadvantage.

- What Black-and-white Mary lacks is the capacity to simulate looking at a red thing: the ability to shift aspect from the view *on* uttering the protocol 'that's red' to the view *from* uttering that protocol. She does not merely lack the understanding from within: grant her that understanding and she might still fail to recognize that a certain physical description goes with looking at red rather than looking at green (Chalmers). She would then lack (and does now lack) the ability to *translate* from the understanding from without to the understanding within, to make the right aspect shift. Plausibly she also

does not know how to perform this aspect shift (Stanley). But know-how is not propositional knowledge. David Lewis's 'ability hypothesis' is not far from the mark. It is not completely on the mark, because there are genuine beliefs available to the perspective from within. But nor is it completely off the mark, because the content of such beliefs is trivial.

- The problem of other minds goes away. There is a certain amount of objective information about someone that would compel me, by force of rational necessity, to enter into sympathy with them. If entering into sympathy with the other settles the question whether they have a mind, the proper amount of objective information is incompatible with skepticism about their 'interiority'.

- A proposal with some similarities to mine (due to John Perry) is that the distinctive first-person understanding of consciousness is 'indexical'. The rough idea appeals to 'centered worlds' theory: belief contents are not propositions but properties (aka sets of possible time-slices aka possible worlds with distinguished centers). The view from within looking at a red thing is captured in the set of centered worlds centered on utterances of the protocol 'that's red'.

  Chalmers objects to this view as follows:

  > In the indexical case, any epistemic gaps disappear from an objective perspective. Say that I am physically omniscient, but do not know whether I am in the USA or Australia (let's imagine that there are appropriate qualitative twins in both). Then I have a certain indexical ignorance, and discovering that I am in the USA will constitute new knowledge. But if someone else is watching from the third-person point of view and is also physically omniscient, they will have no corresponding ignorance: they will know that A is in Australia and that B is in the US, and that's that. There is no potential knowledge that they lack: from their perspective, they know everything there is to know about my situation. So my ignorance is *essentially* indexical, and evaporates from the objective viewpoint. ...
  >
  > Now consider Mary's ignorance. From her black-and-white room, she is ignorant of ... what it will be like for her to see red for the first time ... . [A] physically omniscient observer may have precisely analogous ignorance: even given his complete physical knowledge, he may have no idea what it will be like for Mary

> to see red for the first time. So this ignorance does not evaporate from the objective viewpoint. ... This suggests very strongly that phenomenal knowledge is not a variety of indexical or demonstrative knowledge at all. Rather, it is a sort of objective knowledge of the world, not essentially tied to any viewpoint.

Chalmers's objection against the centered worlds approach hits its mark. Fortunately, my proposal is distinct: let's see how.

Uncertainty about one's location—am I the guy at the east end or the west end of the forest?—is affiliated with uncertainty de re: are *those* trees at the east or the west end of the forest? This uncertainty in turn emerges from uncertainty de hoc: uncertainty about easterly or westerly character, in regard to the trees that are the target of attention. The corresponding de dicto uncertainty is spread over worlds in which *those very trees* are in the east and worlds in which *those very trees* are in the west. This proposition is externalistically individuated: different trees, different proposition. If I *am* in the east end, my uncertainty is spread over worlds in which the easterly trees are easterly and worlds in which the easterly trees are westerly; if I *am* in the west end, it spreads over worlds in which the westerly trees are easterly and worlds in which the westerly trees are westerly. In the former case there aren't any worlds of the latter sort (I'm reading these as identity sentences); in the latter case there aren't any worlds of the former sort. Either way I am not uncertain about anything relevant; but due to the externalism I present as in an indeterminate credential state. My uncertainty is in a sense 'metacognitive', after the manner of uncertainty whether Hesperus is Phosphorus.

Something similar goes for uncertainty about which organism is the one uttering the sentences underlying these thoughts: the east end organism or the west end organism? We can think of the externalism here as rooted in 'token-reflexivity': a self-ascription de hoc of a 'physical' property selects as its standard of truth the organism producing the ascription. So suppose I am uncertain whether I am the east end guy or the west end guy. If I *am* the east end guy all of my epistemic possibilities are ones in which I am the east end guy; and if I *am* the west end guy all of my epistemic possibilities are ones in which I am the west end guy. The sentences 'I am that east end guy' and 'I am that west end guy' are governed by the following wide scope schema:

- If an uttering body is uniquely F, for that body:

$$\Phi(\mathrm{I}) \dashv\vdash \Phi(\text{that } F).$$

Thanks to this externalistic rational governance, my credential state is indeterminate.

The metacognitive character of my uncertainty, stemming from the token-reflexive referential tie of my sentences to the organism uttering them, cannot be appreciated by anyone else (not without an act of simulation). Sam is a distinct organism from me. If she knows that there is a guy lost in the east end and a guy lost in the west end (and she knows all the details about the guys and the forest and yada yada), nothing relevant remains for her to be uncertain about. By contrast, I know all that stuff but remain ignorant. The difference is that the metacognitive uncertainty I face is unavailable to Sam. It is unavailable to her but available to me because the Hellie-organism is a constituent of our common subject-matter, and she is not the Hellie-organism while I am him. This is captured in the inference rule: since Sam is not even in the forest, the inference schema does not fire in application to her whether instantiated with 'east end guy' or 'west end guy'.

In that sense, my uncertainty does indeed 'evaporate from the objective viewpoint'.

Suppose that Black-and-white Mary has seen red and seen green, but she does not yet know which one goes with her utterance of protocols ρ or γ: plug $\mathsf{TO}^{\mathrm{BWM}}[\mathsf{B}\rho]$ or $\mathsf{TO}^{\mathrm{BWM}}[\mathsf{B}\gamma]$ into the objective end of her simulator and neither case of imagination comes out the subjective end; conversely, plug $\mathsf{B}\rho$ or $\mathsf{B}\gamma$ into the subjective end and neither physical sentence comes out the objective end.

In our view, a subjective sentence φ uttered by X has (or has presuppositions with) content equivalent to the objective sentence $\mathsf{TO}^{X}[\varphi]$ (or 'X utters φ'). Mary's uncertainty is metacognitive, stemming from rational governance of the use-mention relation by the following wide scope rule:

- If in uttering S, X means that φ, then for X:

$$X \text{ utters } S \dashv\vdash \varphi.$$

Comments:

1. The following wide scope rule has broader applicability:
   - If in uttering S, X means that φ:

$$X \text{ utters } S \dashv\vdash \mathsf{FROM}^{X}[\varphi].$$

2. Our notation $\mathsf{TO}^X[\varphi]$ keys its meaning to 'X utters S' where by uttering S, X means φ. It is more transparently equivalent to φ, for one who knows oneself to be X.
3. One respect in which Black-and-White Mary's uncertainty does not evaporate from the objective point of view is that the latter wide scope rule is not individual bound. Everyone (everyone with the right sort of simulator?) is bound by that latter rule.
4. The triggering condition for these wide scope rules is nonpropositional. One's grasp of it consists in one's making of the shift of aspect between 'X utters S' and 'X means φ'; its truth consists in nothing further than that 'X utters S' and 'X means φ' present the view from without and from within a single state of the world: that the shift of aspect one makes in grasping it is the correct one to make.
5. This sort of aspect shift does not present a single coherent picture of the world. There is no more inclusive view from which the view from within and the view from without can be incorporated. They exclude one another. Use crowds out mention. That is the lesson of the transparency of consciousness. (Compare Kit Fine's views on 'fragmentalism': each moment presents the transition between the past and the future in a way incompatible with its presentation at any other moment and incompatible with the eternal view.)

   If not, the semantic oddity of the triggering condition runs deeper than its being merely nonpropositional (e.g., after the manner of a test sentence): it is not even merely rationally equivalent to any sentence with propositional content. What it says can appear neither as the object nor the form of consciousness. It concerns the locus at which use and mention are joined; but that is a position that cannot be occupied.
6. In this sense also Black-and-White Mary's uncertainty does not evaporate from the point of view of total objective knowledge. To eliminate the uncertainty one must take the leap. But taking a leap is not the same as learning anything.

- What of the hard problem? Why *does* technicolor phenomenology arise out of soggy grey matter? And is this question uniquely hard?

  It is uniquely hard. This is because of the particular character of the aspect shift involved between the explanadum and the explanans. The view from without encompasses all that which can be viewed from without; my nighttime view on a planet and my daytime view on it can both be incorporated

within a more inclusive total view from without, along with the microphysical characterization of the universe and the laws governing it. If this incorporation within a more inclusive view is what the sort of explanation at issue involves, an explanation of why Hesperus and Phosphorus are both planets can be given.

But as we have seen, the views from within and from without do not fit together into a single coherent picture of the world. The problem of consciousness cannot be solved.

- Chalmers challenges the proponent of the physicalist wide scoping story to provide a 'physical explanation' of our epistemic predicament in regard to consciousness. If I am correct here, this would be to extract acquaintance from the view from without. But, as we have seen, this can't be done.

  Ultimately, each of us must regard our ability to shift perspective between use and mention—and, therefore, the very possibility of our existence within the objective world—as an unfathomable mystery.