Semantic possibility*

Wolfgang Schwarz
Draft, 25 September 2015

1 Semantics and meta-semantics

In ethics, it is common to distinguish between questions of first-order ethics – say, whether torturing kittens is wrong – and question of meta-ethics, about what makes it the case that something is right or wrong and about how we could come to know such facts. A parallel distinction is sometimes drawn in semantics. The task of first-order semantics, on this picture, is to study what linguistic constructions mean, while meta-semantics investigates how meaning facts are grounded in other facts and what kind of evidence is relevant to hypotheses about meaning.

But there is a noteworthy difference here between ethics and semantics. We can be confident that torturing kittens is wrong without first settling or taking a stance on the basic questions of meta-ethics. By contrast, semantics cannot be pursued in isolation of meta-semantics. Is the meaning of 'London' a city, an idea, a property, a function, or a node in a representational structure? The question cannot sensibly be addressed without first clarifying what the assignment of meaning is supposed to do.

The reason for this difference between ethics and semantics is that, on the one hand, a lot of people care deeply about what is right and wrong, in the ordinary sense of 'right' and 'wrong'; the answers have direct and far-reaching consequences on our lives. On the other hand, it is doubtful that anything of theoretical or practical importance hangs on what 'London' means, in the ordinary sense of 'means'. Indeed, it is doubtful that there even is a sufficiently precise and unequivocal ordinary sense of 'meaning' that could be made a subject of inquiry. Even if there were, it is hard to see why a whole academic discipline should be devoted to studying that sense. The central terms of semantics – 'meaning', 'reference', 'semantic value', etc. – are better understood as theoretical terms, more like 'gene' in biology or 'rest mass' in physics than 'right' and 'wrong' in ethics. Ordinary judgements about meaning are relevant because they point towards the phenomena that semantic theories are supposed to capture, not because they constitute the subject matter of semantics.

So what is the subject matter of semantics? What are the relevant phenomena towards which our semantic judgments point? Here opinions vary widely. Some hold that the task of semantics is to provide a model theory, in the logician's sense, for entailment facts.

* Thanks to David Ripley and Clas Weber for comments on an earlier version.

Others see the goal in uncovering the internal workings of our language faculty. Yet others look for an association between utterances and propositions that fits in a broadly Gricean account of communication and speaker-meaning. And so on. These different views in meta-semantics impose very different success conditions on first-order semantics. An adequate semantics for one project may be completely inadequate for another. A computational theory of language processing plausibly calls for a very different kind of semantics than a systematization of entailment relations.

It would be wrong-headed to debate which of these different research projects is the true project of semantics. Semantics is whatever we make it, and it can be many things at once. What one can reasonably debate is which of the semantic projects is worth pursuing.

The interdependence of semantics and meta-semantics goes both ways. One can't do semantics without taking a stance on meta-semantics, but one arguably also can't develop an adequate meta-semantics without paying attention to first-order semantics, to keep track of whether the envisaged project can actually be carried out. The common practice of starting with a vague and sketchy idea of the goal and adjusting it in the course of enquiry makes perfect sense. We don't have to start from first principles.

Nonetheless, it may be useful to occasionally step back and reflect on the overall project, especially since quite different projects are currently pursued under the heading of semantics. Misunderstandings are common, when members of school A complain that the semantic values postulated by school B are not suitable to play the role of meanings envisaged in school A. Worse, some prominent strands of semantic theorizing arguably rest on meta-semantic assumptions that do not withstand scrutiny. I will give an example in section 3.

In the remainder of the paper, starting with section 4, I will try to explicate a particular take on meta-semantics that I think provides an attractive background for the kind of formal semantics developed in the tradition of Montague, Lewis, and Stalnaker. (The conception I propose is not meant to be original, but I will not venture to guess to what extent it matches what particular authors working in that tradition have had in mind.) Semantic values are here construed as abstract mathematical entities somehow involving possible worlds. These entities are clearly not things competent speakers "grasp", in any substantive sense of the term. They are too coarse-grained to capture intuitive differences in meaning or cognitive significance. They are presumably not involved in the computational processes that underlie our linguistic competence. They are meant to play a different kind of role, but it is not always clear what that role is. Historically, they were often used to give a compositional semantics of intensional (especially modal) constructions, but in the project I will outline this is an inessential part of their job description. I will also emphasize that the "possible worlds" of semantics should not be identified with the possible worlds of metaphysics. They should not be taken as

independently given at all, but rather treated as theoretical postulates.

Before I turn to all this, let me begin with some very general remarks on how I plan to approach the subject matter of semantics.

2 Semantics as a special science

Most sciences are "special sciences", dealing with phenomena that are in some sense determined by lower-level physical facts but not explicitly described by physics. Consider ecology. If a population increases or decreases, this is always because individual members of the population migrate or die or reproduce, and ultimately because various physical particles move in particular ways. A population can't increase if the underlying physical state remains the same. Nevertheless, there is a point of studying ecosystems and populations directly, without modelling them as complex swarms of microphysical particles. The reason is that there are robust, systematic patterns in the dynamics of populations – patterns that are largely independent of the microphysical details and that only become apparent once we abstract away from the history of particles and individuals. These patterns are captured by population models.

Semantics, too, is a special science. On any reasonable conception of meaning, the meaning of our words is fixed by lower-level, non-semantic facts about the world. Semantics must therefore earn its keep because there are systematic patterns in those facts that only become apparent when we abstract away from lower-level details.

From this perspective, a job description for semantics should clarify what kinds of patterns are supposed to be captured by assigning meanings to expression. The answer must not appeal to facts about reference or entailment: we are looking for lower-level facts that make it useful to speak of reference and entailment in the first place. To be sure, there are *judgements* about reference and entailment. That is, there is the fact that people are disposed to assent to sentences of the form 'A entails B', or 'X refers to Y'. But it would be disappointing if all that semantics could offer is a systematization of this very narrow range of verbal behaviour.

So what kind of non-semantic events in the world might be systematized by introducing semantics? Here is a promising example. Alice is on her way to a coffee shop when she meets Bob. During their brief encounter, Bob utters the sounds 'the coffee shop is closed'. As a consequence, Alice does not continue towards the coffee shop but instead turns towards the university cafeteria.

This interaction could be analyzed on many levels. As with populations, we could study Alice and Bob as complex swarms of microphysical particles. But there is also an intentional and semantic account of what happened. It goes roughly as follows. Alice wanted to get a coffee from the coffee shop. Bob then uttered a sentence meaning that the coffee shop is closed. Understanding and trusting his assertion, Alice came to believe

that the coffee shop is closed, and consequently gave up her plan to get a coffee from there.

This is a high-level description insofar as it is neutral on physical and chemical details. The description fits not just the actual interaction between Alice and Bob, but also many counterfactual variations of that interaction, even including interactions in which Bob uttered different sounds. It captures a pattern that is common to all those interactions. To a first approximation, the pattern in question concerns the way in which Alice's behaviour is affected by Bob's utterance. And of course this is a pattern we don't just find in exchanges between Alice and Bob. If Bob had met Carl rather than Alice going to the coffee shop, we would expect Carl to respond in a similar way to an utterance of 'the coffee shop is closed'.

Generalizing, we might propose a rudimentary model of linguistic behaviour along the following lines. First, we postulate an assignment of meanings to sentences: 'the coffee shop is closed' means that the coffee shop is closed, and so on. Second, we postulate general principles about the production and reception of sentences with a given meaning. Thus: when a member of a linguistic community hears a sentence S which means that p, they tend to believe that p and consequently act in ways that are reasonable given p.

As it stands, this model is way too sketchy to be of much use. But it is sufficient to illustrate what a high-level model of language use might look like. What makes such a model adequate is not that it gets the fundamental meaning facts right. There are no fundamental meaning facts. What an adequate model needs to get right are certain patterns in the production and reception of words in the relevant community. The assignment of meanings in the first part of the model has empirical content only through the second part. One cannot directly observe whether 'the coffee shop is closed' means that the coffee shop is closed. But one can observe whether people have a tendency to respond to utterances of that sentence in ways that would be reasonable if the coffee shop is closed.

One might worry that the postulated meanings are somehow redundant or epiphenomenal. When Alice changes her course in response to Bob's utterance, a lot is happening in her nervous system: neurons fire, glucose is metabolized, neurotransmitters are released. It is those events that ultimately make her turn to the cafeteria. The supposed meaning of Bob's utterance does not show up here. If a neuron fires, that is always in response to chemical activation. Abstract meanings seem to be causally inert: they cannot push or pull things around.

This kind of worry has been voiced in almost every special science. If every event in the world can be predicted as the result of interactions between microphysical particles, how can there be any room for genes or populations, let alone more abstract components of high-level models such as the binomial probability measures in theories of genetic drift? These things play no role in the interactions between microphysical particles. So what is

the point of postulating the genes or probability measures? Shouldn't we rather focus on the lower-level things that do the actual pushing and pulling?

We should not. For one thing, on philosophical reflection it is not at all clear that the relevant abstract entities are causally inert. If causation is understood in the popular tradition of counterfactual or interventionist accounts, then the values taken by abstract probability measures or meaning functions may well come out as genuine causes. Even setting this point aside and granting that meaning functions or probability measures are causally inert, it would be a mistake to ban them from scientific models. Remember that the point of models in special science is to capture interesting and robust patterns in the world that somehow emerge from the chaotic complexity of microphysical interactions. If abstract parameters help achieve this goal, that is enough to vindicate their use.

In linguistics, the worries about epiphenomenology often lead to the idea that semantics should be part of computational psychology, studying what goes on in speech production and processing – preferably in a way that does not involve abstract propositions, assignment functions, sets of worlds or the like. The computational psychology of language processing is certainly an exciting field, but I do not think we should reduce all of semantics to this topic. The patterns of language use towards which I gestured above do not turn on computational details. Perhaps in the not-too-distant future we will communicate in English with robots employing very different computational processes. If semantics is about these processes, we will have to say that the robots speak a different language: the semantics of human English will look completely different from the semantics of robot English. But clearly there is another sense in which the robots and us speak the same language. And that has something to do with overt patterns of use: a robot on a mission to get coffee at the coffee shop will also tend to change its plan in response to utterances of 'the coffee shop is closed'. There are higher-level patterns here that are worth studying, independent of their computational or microphysical realization.

In fact, even if we set ourselves the task to study language processing, we arguably need guidance from such higher-level semantic theories. Before we can find out how the brain implements a certain ability, we have to know what we are looking for, what the relevant ability amounts to. It would not be illuminating to learn that such-and-such computational events happen in the processing of 'the coffee shop is closed' unless we get at least some insight into how these processes constitute an understanding of the utterance. And arguably understanding largely consists in displaying the right kind of psychological and behavioural response to the relevant token. If you understand a token of 'the coffee shop is closed', you will come to believe (assuming the speaker is trustworthy and knowledgeable) that the coffee shop is closed, with corresponding effects on your behaviour.

3 Propositionalism

I have outlined a very basic model of communication. In the present section, I want to explain how we should *not* flesh out that model.

First some more platitudes about the encounter between Alice and Bob. Before their interaction, Bob had the belief that the coffee shop is closed. He uttered 'the coffee shop is closed' with the intention that Alice, too, should come to believe that the coffee shop is closed. By uttering the sentence, Bob expressed his belief that the coffee shop is closed. He asserted that the coffee shop is closed. Alice grasped what Bob asserted: that the coffee shop is closed. Trusting his assertion, she came to believe that the coffee shop is closed.

To philosophers who are trained to think in the categories of objects, properties, and relations, these platitudes might suggest that there is an object denoted by the clause 'that the coffee shop is closed', which is initially believed by Bob ('Bob believes that the coffee shop is closed'), then asserted through his utterance ('Bob asserts that the coffee shop is closed'), and subsequently believed by Alice ('Alice believes that the coffee shop is closed'). The object also seems to have a truth value ('That the coffee shop is closed is true'), it has modal properties ('That the coffee shop is closed is possible'), and it stands in logical relationships to other such entities ('That the coffee shop is closed entails that there is a coffee shop').

Such objects are known as *propositions*. We might construct a model of communication in terms of propositions. At each time, speakers and hearers stand in the belief relation to various propositions. If a speaker believes a proposition p and wants a hearer to share that belief, she can utter a sentence which means that p. The process might be fleshed out along Gricean lines: in a paradigmatic case of communication, the speaker intends to produce in the hearer a belief that p by making her recognize this very intention, etc. (see [Grice 1957]).

Assuming a model along those lines, we could take the basic task of semantics to be the definition of rules that determine what proposition is expressed by the utterance of any given sentence in any given utterance context. We can expect there to be a systematic story here, since the number of sentences we can use for communication is unbounded. There should be some kind of general, recursive recipe that determines the propositions expressed by sentences in contexts. This recipe is the semantics of the relevant language.

The complete model still resembles our first stab from the previous section: the semantic rules that determine an association of sentences with propositions will be accompanied by general principles concerning the production and reception of sentences. These principles might say, for example, that people who utter a sentence S which means that p generally intend their audience to believe that p. The model is adequate if it gets the facts about beliefs, intentions, etc. right. In a further step, one might then hope to reduce those facts

about beliefs and intentions to lower-level, non-intentional facts.

This approach, which I will call 'propositionalist', has some intuitive appeal and used to be quite popular among philosophers, especially in the 1980s. Nevertheless, I think it should be rejected, for two main reasons.

The first is that the propositionalist analysis of attitude reports and speech act reports is plausibly inadequate: such reports do not simply express a relation between the subject and some kind of entity denoted by the embedded *that*-clause. Second, even if ordinary attitude reports could be analysed in terms of relations to propositions, it is doubtful that these relations could form a suitable basis for a science of meaning.

Let me begin with the first problem. Suppose Carl believes that Bob is overseas while Bob is hiding in the closet. Then 'Carl believes that the man in the closet is overseas' can be true, even though Carl presumably does not stand in the belief relation to the (obviously false) proposition expressed by 'the man in the closet is overseas'. Indeed, there is also a reading on which 'Carl believes that the man in the closet is overseas' is false, and we should not infer that Carl both believes and disbelieves one and the same proposition. So 'Carl believes that p' does not simply attribute to Carl a belief in the proposition expressed by 'p'. Kripke's [1979] "puzzle about belief" indicates that the present problem is not limited to a well-circumscribed class of "de re" reports. Recall also Mates's [1950] observation that attitude reports can be sensitive to orthographic and phonetic features of the embedded sentence that plausibly make no difference to the expressed proposition: if Alice has erroneous views about the meaning of 'ophthalmologist', then 'Alice believes that ophthalmologists are ear doctors' can be true while 'Alice believes that eye doctors are ear doctors' is false even though the embedded sentences presumably express the same proposition. (If they don't, the danger is that almost no two sentences express the same proposition, contradicting the fact that we often judge that two people said the same thing even though they didn't use the exact same words.)

Moreover, even if the simple relational analysis of belief sentences were tenable, other embeddings of *that*-clauses arguably can't be analysed as relations to the very same kind of entities. Consider proving-that. The objects of a formal proof are highly structured entities: in some areas of maths and logic, it can make a big difference whether one has proved $p \wedge q$ or $q \wedge p$. By contrast, visual perception plausibly isn't directed at syntactically structured entities. There is no difference between seeing that a given ball is red and round, and seeing that it is round and red. So a *that*-clause embedded under 'seeing' appears to denote a different kind of entity than the same *that*-clause embedded under 'proving'. A much discussed instance of this phenomenon concerns temporal embeddings (see e.g. [Weber 2012]): in a statement such as 'tomorrow, it will be the case that the cat is on the mat', the complement clause seems to pick out a "temporally neutral proposition" that takes different truth values at different times. By contrast, embedded in 'Bob believes that –' or 'it is possible that –', the same clause is

generally taken to denote a temporally non-neutral ("eternal") proposition.

Another problem for the propositionalist analysis is the variability and context-dependence of ordinary attitude and speech act reports. If Alice knows that Bob is an eye doctor but mistakenly believes that 'ophthalmologist' means ear doctor, does Alice know (or believe) that Bob is an ophthalmologist? Does she *say* that Bob is an ophthalmologist when she says that Bob is an eye doctor? The answer arguably varies with the context of attribution. If we let our semantics be guided by ordinary attitude and speech act reports, we would have to conclude that the semantic value of every utterance is relative not just to the context of utterance but also to a third-person context of attribution.

Here we see how first-order semantics can be important to meta-semantics. The evidence from first-order semantics strongly suggests that English statements involving *that*-clauses cannot be analysed as simple predications involving special kinds of entities unambiguously denoted by *that*-clauses. If propositions are defined as those entities, the evidence suggests that propositions do not exist.

Now let's set all these problems aside and pretend that we can analyze ordinary attitude reports and speech act reports as stating relations between subjects and propositions. It is a further question whether those relations should play a central role in our meta-semantics. From the perspective of the previous section, the idea should look suspicious. The goal of semantics as a special science is not to systematize or predict English judgements about beliefs and intentions, or about "what is said" by an utterance. The goal is to capture interesting patterns in lower-level, non-semantic and non-intentional events. But there is no good reason why a systematic model of this kind should closely follow our ordinary way of talking about belief, assertion, and communication. Indeed, there are several reasons against this assumption.

First of all, systematisations of ordinary attitude reports are beset by a range of thorny technical problems. For example, suppose we formalize knowledge claims by means of a binary predicate K relating a subject s to a proposition p; define K^* as the logical closure of K, so that $K^*(s,p)$ holds whenever p logically follows from some q such that $K(s,q)$. Since knowledge entails truth and logical consequence is truth-preserving, all instances of $K^*(s,p) \supset p$ should be true. If the agent s is aware of this fact, then $K^*(K^*(s,p) \supset p)$ should be true as well. Yet Montague [1963] proved that if there is a recursive mapping from sentences to propositions then these apparently harmless assumptions are inconsistent with basic arithmetic.

Second, the truth-conditions for ordinary-language attitude reports plausibly depend on semantic facts about the meanings of words and sentences. What is reported by attitude reports then aren't suitably lower-level facts. On the propositionalist approach, semantics, together with the hypothesized propositional-attitude psychology, postulates a network of interconnected relations to propositions, but what we would like to know

is how that network is grounded in non-intentional facts: What underlying patterns are captured by postulating the network? There is a striking lack of progress on this question.

Third, it has been argued that while ordinary attitude reports are "wide" insofar as the truth-value of such reports is often sensitive to the identity of external objects in the subject's environment (and even to linguistic conventions in the subject's community), the proper analysis of such reports turns on a deeper conception of belief that is "narrow". The basic idea is that, say, 'Ralph believes that Ortcutt is a spy' reports that (a) Ortcutt is known to Ralph as the occupant of some role R and (b) according to Ortcutt's belief state, whoever occupies R is a spy. Independently of these proposals there is a long-standing debate in philosophy and cognitive science over whether a narrow conception of belief should be used for the purposes of a systematic study of mind and cognition. The propositionalist approach presupposes that the answer is no.

Fourth, as a matter of fact the most powerful models of belief, learning, communication, and rational action developed in recent decades do not follow the propositionalist model. Bayesian accounts of perception, belief, confirmation, and rational decision employ subjective probability measures that do not simply mirror ordinary-language attitude reports. The objects in the domain of the probability measure are certainly ill-suited to play the role of propositions in the propositionalist picture. The same is true for the sets of worlds used in contemporary models of binary belief and communication, as studied for example in the logic of action and public announcement ([van Benthem 2011]).

It may be worth highlighting one of the reasons why the propositionalist model has not proved terribly useful. The problem is to explain how propositional attitudes manifest themselves in behaviour. For concreteness, suppose we construe propositions as *sui generis* abstract Fregean "thoughts". On the propositionalist account, when Alice heard Bob say 'the coffee shop is closed', she came to stand in the belief relation to one of those entities. Why did that make her change her course? Why would she have behaved differently if she had stood in the belief relation to other Fregean thoughts? There must be some connection between the thoughts or propositions on the one hand and types of behaviour on the other. A propositionalist might say that by turning to the cafeteria, Alice *made true* the proposition that she will get coffee, which is something she desires. But that only puts a label on the puzzle. How does a *sui generis* abstract entity come to be "true" by going to a cafeteria, and why does Alice care about whether a certain abstract entity is "true"?

I don't claim that the propositionalist approach is refuted by these problems. Large-scale rescue attempts have been launched to tackle some of the issues I have listed. Nonetheless, I think it would be advisable not to bet one's approach to semantics on the outcome of these developments. In any case, I will outline a different approach.

4 Possible-world semantics in signalling models

How should we model core aspects of language use without presupposing the propositionalist model? An illuminating starting point are signalling models.

Consider the following type of scenario. There are two agents, which we'll call 'the sender' and 'the receiver'. The sender is in a position to observe a certain state of the world – the presence or absence of a predator, say. The receiver has a choice between a number of acts, and it would be advantageous to the community of sender and receiver if her act were sensitive to the state observed by the sender, so that she chooses act A_1 if the state is W_1, A_2 if the state is W_2, etc. However, the receiver cannot directly observe the relevant state. Fortunately, the sender also has at her disposal a range of acts, called 'signals', that can be observed by the receiver and that don't have any direct advantages or disadvantages in themselves. If the sender chooses different signals depending on the observed state of the world, the receiver can make her acts depend on the state by making her acts depend on the observed signal. In order to achieve the advantageous correlation between receiver acts and states, the sender and receiver must then coordinate on a signalling strategy. If the sender produces signal S_1 upon observing state W_1, but the receiver responds to S_1 by choosing act A_2, the goal of responding to W_1 with A_1 is not achieved. To secure the advantageous dependence of receiver acts on the observed state, there must be a mapping M from signals to states such that (a) the sender reliably produces signal S_i only if she observes the state $M(S_i)$, and (b) the receiver reliably responds to S_i by choosing the act appropriate for $M(S_i)$.

Signalling is pervasive in the biological world. Many species of monkeys, birds, bees, and even bacteria display systematic responses to certain environmental stimuli whose evolutionary purpose is to allow other members of their group to respond in an adequate way to these stimuli, which they may not have received themselves. There is an extensive literature on the evolution of signalling systems and on how much light this may shed on the evolution of human language (see e.g. [Skyrms 2010], [Stegmann 2013]). I do not want to get into these issues here. The reason why I bring up signalling models is that they provide a helpful illustration of possible-worlds semantics.

A signalling model of the kind I described includes an association M between signals and states of the world. In most situations, the majority of these states are merely possible: a given signal may be associated with the presence of a leopard even if in fact no leopard is present. So the "meanings" assigned to signals are possible states of the world. Such states can be more or less specific. Suppose there is both a signal for the presence of a leopard and another signal for the presence of a large animal. Every state in which a leopard is present is also a state in which a large animal is present. The standard way to capture such facts is to model states as sets of maximally specific states – i.e., possible worlds. Thus a signal for leopards will be associated with the set of possible

worlds in which a leopard is present.

Evidently, the point of using possible worlds (or world states) in signalling models is not to analyse the meaning of modal expressions; signalling languages usually don't have any modal expressions. Nor is the point to systematize judgements about truth-conditions or entailment. The point of assigning a given set of worlds to a given signal is that this allows for elegant and systematically useful way to capture certain behavioural patterns in the relevant community. What is captured by a mapping M from signals to states of the world in a signalling model is a counterfactually robust regularity in the production and response to the relevant signals: senders generally produce a signal S_i only in state $M(S_i)$, receivers generally respond to S_i in a manner that is advantageous in $M(S_i)$. This is how a semantic hypothesis concerning the association between signals and sets of worlds has empirical implications, despite the fact that possible worlds do not figure in a low-level account of the processes that take place when a sender produces a signal or a receiver responds.

My sketch of signalling models needs some refinements. For example, we might want to model certain alarm calls as signals for the presence of predators even if most actual calls are false alarms. A natural thought here is to make use of the fact that the call's evolutionary function is to be produced in the presence of predators; see [Adams and Aizawa 2010] and [Neander 2012] for some attempts at spelling out this idea. I won't descend further into this difficult issue here.

When we turn to human language use, we have to go beyond simple signalling models anyway. We need to connect the production and reception of utterances with the beliefs, expectations, rights, and obligations of the relevant individuals. We also need to take into account the productivity and creativity of human language use. And obviously human languages are not innate, so evolutionary considerations are largely irrelevant to the semantics of Latin or English.

On the other hand, we should not exaggerate the differences. Some aspects of human language use look a lot like signalling. People in the English-speaking community tend to produce 'it is raining' only when it is raining, and they tend to respond to such productions in a rain-adequate manner. Moreover, the evolutionary models for the emergence of signalling systems are easily adapted to models of cultural evolution and individual learning. Instead of completely abandoning the signalling model, we might therefore do well to expand it.

5 Bayesian pragmatics

In order to capture the interplay between utterances and attitudes, we need a systematic model of attitudes. A promising candidate here is the Bayesian model as outlined for example in [Ramsey 1931] or [Jeffrey 1983].

A standard Bayesian model associates with every agent i at every time a pair of a probability measure P_i and a desirability measure D_i over possible states of the world. Intuitively, the P_i value assigned to a state reflects the agent's degree of confidence that the state obtains, while the D_i value reflects the extent to which the agent would like the state to obtain. However, we do not assume that these measures track familiar introspectible quantities, nor do we assume any simple connection between the measures and attitude reports in ordinary English.

Instead, what the assignment of probabilities and desirabilities is meant to capture is a certain causal-functional profile of the agent's cognitive state. A helpful and popular analogy here are physical measures such as mass or temperature (see e.g. [Stalnaker 1984: ch.1]). In order to have a temperature of 153°C an object does not need to stand in some metaphysically substantive relation to the abstract number 153. The assignment of numerical temperatures merely provides a useful way to pick out certain (rather complicated) functional properties of physical objects.

In the case of probabilities and desirabilities, the relevant functional profile concerns first and foremost the agent's behavioural dispositions. According to Bayesian decision theory, an option in a choice situation is licensed by given probability measure P_i and desirability measure D_i just in case it maximizes the expected value of D_i relative to P_i. Thus we might say that an attribution of probabilities and desirabilities in a Bayesian model of certain agents is correct only if the agents' actual and counterfactual choices are (*ceteris paribus*) licensed by the attributed probabilities and desirabilities. Another important component in the functional profile of probabilities and desirabilities concerns the way in which these measures evolve in response to sensory information. On the simplest version of the story, P_i evolves by conditionalization, so that the new value of $P_i(A)$ equals the old value of $P_i(A/E)$, where E captures the new information, and D_i evolves so that the new value of $D_i(A)$ is the old value of $\sum_{w \in A} D_i(w) P_i(w/AE)$.

This simple model obviously ignores all kinds of subtleties in human psychology. More complicated versions of the Bayesian model have been proposed, but the complications are not terribly important for our present topic, so let us move on to language.

Utterances systematically affect subjective probabilities. When people who understand English hear an utterance of 'the coffee shop is closed', they tend to assign relatively high probability to situations in which some contextually salient coffee shop is closed, and thus tend to act in a way that makes sense on the assumption that the coffee shop is closed.

The posterior probability of H after learning E is high iff the prior conditional probability of H given E is high. Thus before Bob said 'the coffee shop is closed', Alice already gave high probability to situations in which the coffee shop is closed conditional on hearing Bob utter 'the coffee shop is closed'. More colloquially, Alice expected Bob to utter those words only if the coffee shop is closed.

Moreover, this expectation is presumably known by Bob, in the sense that Bob assigns high probability to states in which Alice has the expectation. This is part of the reason why uttering 'the coffee shop is closed' maximizes expected desirability for Bob: Bob utters 'the coffee shop is closed' because he gives comparatively high desirability to states in which Alice avoids going to the coffee shop in vain. According to Bob's subjective probability measure there is a good chance that in response to his utterance Alice will assign high probability to states in which the coffee shop is closed, which will prompt her not to go there in vain.

Given these observations, one might try to define an interpretation function M from utterance types to possible states of the world such that people in the relevant community generally (*ceteris paribus*) assign high conditional probability to $M(S)$ given an utterance of S, as well as to states in which other members of the community assign high conditional probability to $M(S)$ given an utterance of S. More colloquially, people expect specific utterances to be produced only if the world is a certain way, and they expect others to share those expectation.

However, the resulting model would be very cumbersome, and its ill-behaved interpretation function would bear little resemblance to any pre-theoretic conception of truth-conditions or meaning. For example, on the present approach the states associated with 'I am tired' might be states in which the speaker will soon go to sleep, those associated with 'there is a bird in the garden' will exclude states in which the relevant bird is a penguin, and those associated with 'the horse raced past the barn fell' will be ones in which the speaker is making some sort of linguistic point.

Intuitively, the problem is that statements with a certain literal meaning are commonly used to convey more (or even altogether different things) than their literal meaning. Introducing something like this distinction plausibly allows us to greatly simplify the semantic part of our model. The idea is to factorize the present interpretation function M into a "literal semantic" component $L : S \to 2^W$ and a "pragmatic" component $P : S \times 2^W \to 2^W$ that makes use of the output of the L function. For example, the semantic function might assign to 'some students passed' the set of all states in which at least one of the contextually salient students passed, and the pragmatic component might exclude from this set the states in which every student passed. Following [Grice 1989], one might further hope that the pragmatic function P can be explained by language-independent, general principles of rationality and cooperativeness, but it is an open question to what extent this is possible.

(The distinction between L and P might be further reflected in certain aspects of language use, such as to the possibility of "cancellations". More directly, we arguably tend to criticise people differently for saying something false than for saying something misleading. Grice also suggested that literal content is processed first, but I don't think we should put much weight on that conjecture.)

Much recent work in semantics, as described for example in [Heim and Kratzer 1998] or [Jacobson 2014], can be interpreted as systematizing the intermediary function L of such a model. These works typically present rules that specify how the literal truth or falsehood of sentences is determined relative to a context. In other words, they describe a mapping from sentence types to functions from possible states of the world ("contexts") to truth-values.

Since humans fairly quickly acquire the right attitudes towards the production and reception of a large number of possible utterances, there must be systematic, recursive rules that determine the functions L and P. Some of these rules plausibly concern the syntactic parsing of sentences, others concern the semantic interpretation of the parsed sentences.

The present approach is completely neutral on what those rules might look like. In particular, it does not imply that when a sentence A is embedded in another sentence B, then the truth-value of B is a function of the set of worlds associated (by L or M) with A. It certainly does not make the absurd prediction that whenever 'Alice believes that A' is true, and B is logically equivalent with A, then 'Alice believes that B' is true as well. The sets of worlds associated with sentences are not supposed to play the role of "propositions" in the propositionalist picture, and they are not supposed to be "compositional semantic values" in the sense that they contain everything on which the truth-value of complex expressions might depend. If our preferred semantics makes use of compositional semantic values – which should not be taken for granted – those values might well involve further 'index coordinates' ([Lewis 1980]) such as assignment functions ([Rabern 2012]), infinite sequences of possible worlds and times ([Cresswell 1990]), information states ([Yalcin 2007]), and what not.

The model I have outlined is deliberately vague and incomplete. I want to advertise a general approach, not a particular theory. The approach allows for many refinements, extensions and variations. As for variations, we could take the meaning function(s) to output not a set of possible worlds, but a two-dimensional function from possible worlds (or "contexts") to sets of worlds, as in [Kaplan 1989]. The usual motivation for this move is that the context-relative sets of worlds are better suited to play the intuitive role of propositions, but it might also turn out that certain actual aspects of our language use are better captured with a two-dimensional semantics. Relatedly, we could hold that the output of the intermediate meaning function L is some kind of "minimal proposition" that only determines a set of possible worlds when enriched by various features of an utterance context. We could also model the impact of utterances in terms of updates on a Stalnakerian [2002] "common ground" that is only indirectly related to the probabilities and desirabilities among the participants to a conversation. And so on.

6 Semantic space

I have argued that possible worlds have an important role to play in systematic models of language use, just as they do in models of signalling among animals. There is nothing special here about semantics. Science is full of probabilistic models whose "sample space" involves mere possibilities.

Accepting such models does not mean to believe that possible worlds somehow figure in the fundamental fabric of reality. It does not contradict the assumption that reality is at bottom completely physical. Physicalism and naturalism plausibly preclude the assumption that there are primitive facts of meaning or representation. They do not require that we only quantify over things explicitly recognized by fundamental physics. That would be the end of special science. On the model-based account I have outlined, the semantic facts are meant to capture high-level abstract patterns in the physical world: if a certain signal can be usefully modelled as representing such-and-such a state of the world, this is always because of lower-level facts about how the signal is used. The representation facts are grounded in ultimately non-intentional facts about usage. That is all the naturalism we should ask for.

Appealing to abstract entities in theoretical models is common practice and harmless. However, if we postulate probability measures or meaning functions, we have to clarify what values these functions can take and what lower-level facts are reflected by those values.

Philosophers often associate "possible worlds" with the idea of "metaphysical modality". The usual story is that there is a special class of judgements about what could or could not have been the case that can be analyzed by appeal to a world-relative assignment of truth-values to propositions; for example, 'it could have been the case that grass is red' is taken to attribute metaphysical possibility to the proposition that grass is red – a status which is then further analysed as truth at some possible world.

It should be clear that the approach to semantics I have outlined in the previous section has very little connection to this story. To be sure, it might turn out that the space of possible worlds assumed in the philosophical analysis of metaphysical modality can do double duty as the space of possible worlds in pragmatic models of language use – or, say, as the sample space for probabilistic models in ecology. But on the face of it, the two job descriptions look very different.

Another, loosely related, approach to possible worlds begins with the observation that propositions can be conjoined, disjoined, and negated. Under certain (non-trivial) assumptions, the set of propositions together with these operations satisfies the conditions on a complete, atomic Boolean algebra. The possible worlds are then identified with the atoms in this algebra. By Stone's representation theorem, the propositions are isomorphic to the sets of possible worlds.

This approach assumes that we have a well-defined space of "propositions" to begin with. But we don't. In previous sections, I have informally appealed to "possible states of the world", giving examples such as "the presence of a leopard", but I don't think our pre-theoretic conception of possible states of the world is very clear. We should take care not to fall back into the propositionalist mistake of assuming that a well-defined space of propositions is involved in our ordinary talk about belief, assertion, possibility, etc. Indeed, the space of "that-clause propositions" is not only ill-defined, but also plausibly fails the algebraic conditions required for the construction of possible worlds as Boolean atoms.

In general, we should not identify the possible worlds in our linguistic models with a fixed set of independently specified entities. The possible worlds are simply part of the model; the construction of logical space is a modelling choice.

So how should we make that choice? A good idea is to start with actual, concrete situations in which an utterance or non-linguistic action takes place. Other possible worlds may then be construed as variations of, or abstractions from, these situations. Remember that the assignment of possible-worlds meanings to sentences is supposed to reflect systematic effects of utterances on subjective probabilities, and that choice behaviour is central to the functional profile that defines subjective probability. When we make a choice, we don't merely select a region in some abstract algebraic space; rather, we bring about some concrete state of affairs. A promising idea is therefore that we assign comparatively high (expected) desirability to certain aspects of that state.

We may then ask what aspects of a concrete state are relevant; in other words, what kind of variations or modifications should we allow for? There are many interesting and difficult issues here.

One question that has received quite a bit of attention is whether we should distinguish between possibilities located in the same possible universe, but at different times or places. There are good reasons, both intuitive and systematic, for a positive answer (see e.g. [Lewis 1979]). Intuitively, objective knowledge about the universe is never enough to know who we are or what time is now. Two agents might have the very same views about the universe but locate themselves at different places. This difference will plausibly show up in different behaviour, and in different responses to evidence. The most straightforward – although not the only – way to capture these phenomena in the Bayesian framework is to make the objects of probability and desirability "centred". The idea has also proved useful in the analysis of context-dependent expressions (where contexts can hardly be modelled as uncentred worlds) and the semantics of various intensional constructions (e.g. [Anand and Nevins 2004], [Santorio 2012]).

Another question familiar at least to philosophers is whether we need to distinguish between possibilities that agree in the distribution of qualitative properties but disagree in which individuals instantiate those properties. Imagine a (centred) scenario in which

a leopard is approaching. Should we distinguish between that situation and an alternative situation in which a qualitatively identical, but numerically different leopard is approaching? The question is not whether there are non-qualitative differences between metaphysically possible worlds; the "metaphysically possible worlds" are not our topic. It is not too hard to construe a logical space in which non-qualitative distinctions can be drawn. The question is whether such distinctions will be needed in our models. Arguably, Bayesian models of learning and rational choice go more smoothly if non-qualitative distinctions are ignored (see e.g. [Chalmers 2011] for the learning side), but the issues are tricky. Note that if we go for purely qualitative worlds, then the semantic contribution of a proper name or singular variable to possible-worlds content cannot be specified by merely picking out an individual in the actual world: there is always a non-trivial question about what it picks out at other possible worlds.

Here is another question. Should we allow for possible worlds that agree in all non-moral respects, but disagree on what is right and wrong? Again, our question is not about metaphysics. Even if the moral supervenes on the non-moral in the space of metaphysically possible worlds, adding a dimension of moral facts might prove useful when it comes to representing people's attitudes towards normative questions, or to model the effect of normative utterances (see e.g. [Gibbard 1990]).

Suppose we decide to allow for purely moral variations among possible worlds. As in the case of non-qualitative variations, the relevant construction does not raise insurmountable technical problems. But suppose we also have the metaphysical view that there are no primitive moral facts. Indeed, suppose there are no substantive moral facts at all. Our logical space then has the curious property that a concrete utterance situation might be represented by a whole range of different possible worlds. For if there are no moral facts, then any concrete situation is exhausted by its non-moral features; so if it is represented by a possible world in our logical space, then it is equally represented by all merely normative variations of that world.

Let's say that a model satisfies *Uniqueness of Actuality* if it does not have this curious property. That is, a model of logical space satisfies Uniqueness of Actuality if every actual situation is represented by at most one world in the model.

If a model fails Uniqueness of Actuality, it gives rise to a kind of relativism. Suppose an agent asserts a sentence whose semantic value is a set of worlds that verify some controversial moral hypothesis – say, that abortion is wrong. Normally, we would say that the assertion is true if the actual utterance context is (represented as) a world in the relevant set. But if the utterance context does not have a moral dimension, then it will correspond to many different worlds in logical space, some of which fall in the set associated with the utterance and some of which do not. So the utterance is neither straightforwardly true nor untrue; it is true only relative to a normative parameter that is not fixed by the utterance context. This might be exploited to explain otherwise puzzling

aspects in the production and reception of normative utterances.

There are many other opportunities for this sort of move. Facts about the future in worlds with branching time are an obvious example, as are the information states relative to which epistemic and deontic modals seem to be evaluated. Or consider conditionals. In many utterance contexts there is arguably no fact of the matter about what is (or would be) the case on various counterfactual suppositions: if there is a golden mountain, how many peaks does it have? On the other hand, it has been argued that introducing primitive conditional facts into logical space allows for a semantics of conditionals that is theoretically elegant (e.g. [Klinedinst 2011]) and vindicates the intuitive idea that the probability of a conditional 'if A then B' generally equals the conditional probability of B given A (e.g. [McGee 1989]).

Another candidate application is to vagueness. The arguments in [Williamson 1994] suggest that vague predicates all have sharp cut-off points. Williamson himself suggests that these cut-offs are fixed in some opaque manner by the utterance context – by subtle details in our usage of the relevant words. But one could also treat the location of the cut-offs as a primitive dimension of semantic space.

One might also try to apply this strategy to hypotheses in mathematics or metaphysics. For example, it might prove useful to distinguish between worlds where mereological universalism, or the continuum hypothesis, is true and others where it is false, even if we (as model builders) do not believe that there is a corresponding objective fact of the matter, so that a given context of utterance or belief does not select one of these possibilities as true or false.

In each application, however, it is not enough to simply postulate the relevant distinctions. One also has to adjust the pragmatics. In particular, we need a story about how probabilities over the alleged possibilities are supposed to manifest themselves in behaviour.

This is not the place to explore these ideas. I only want to scratch the surface in order to show that it is worth digging further. There are different ways of constructing logical space. There is not one true, antecedently fixed space of possible worlds, just as there are no true meanings independently of any semantic model. Meanings and possible worlds are theoretical postulates whose purpose in a semantic model is to capture interesting patterns in an ultimately non-linguistic and non-intentional world.

References

Fred Adams and Ken Aizawa [2010]: "Causal Theories of Mental Content". In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy,* Spring 2010 edition

Pranav Anand and Andrew Nevins [2004]: "Shifty operators in changing contexts". In *Semantics and Linguistic Theory,* 20–37

David Chalmers [2011]: "Frege's puzzle and the objects of credence". *Mind*, 120(479): 587–635

Max J. Cresswell [1990]: *Entities and Indices*. Dordrecht: Kluwer

Allan Gibbard [1990]: *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press

Paul Grice [1957]: "Meaning". *The Philosophical Review*, 66(3): 377–388

— [1989]: *Studies in the Ways of Words*. Cambridge (Mass.): Harvard University Press

Irene Heim and Angelika Kratzer [1998]: *Semantics in Generative Grammar*. Malden (Mass.): Blackwell

Pauline Jacobson [2014]: *Compositional Semantics*. Oxford: Oxford University Press

Richard Jeffrey [1983]: *The Logic of Decision*. Chicago: University of Chicago Press, 2nd edition

David Kaplan [1989]: "Demonstratives". In Joseph Almog, John Perry and Howard Wettstein (Eds.) *Themes from Kaplan,* New York: Oxford University Press, 481–564

Nathan Klinedinst [2011]: "Quantified conditionals and conditional excluded middle". *Journal of Semantics*, 28(1): 149–170

Saul Kripke [1979]: "A Puzzle About Belief". In Avishai Margalit (ed.), *Meaning and Use*, Dordrecht: Reidel

David Lewis [1979]: "Attitudes *De Dicto* and *De Se*". *The Philosophical Review*, 88: 513–543

— [1980]: "Index, Context, and Content". In S. Kanger and S. Ohmann (Eds.) *Philosophy and Grammar,* Dordrecht: Reidel

Benson Mates [1950]: "Synonymity". *University of California Publications in Philosophy*, 25: 201–226

Vann McGee [1989]: "Conditional Probabilities and Compounds of Conditionals". *The Philosophical Review*, 98(4): 485–541

Richard Montague [1963]: "Syntactical Treatments of Modality, with Corollaries on Reflection Principles and Finite Axiomatizability". *Acta Philosophica Fennica*, XVI: 153–167. Reprinted in [Montague 1974: 286–302].

— [1974]: *Formal Philosophy*. New Haven: Yale University Press

Karen Neander [2012]: "Teleological Theories of Mental Content". In Edward N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy,* Spring 2012 edition

Brian Rabern [2012]: "Against the identification of assertoric content with compositional value". *Synthese*, 189(1): 75–96

Frank Ramsey [1931]: "Truth and Probability (1926)". In R.B. Braithwaite (Ed.) *Foundations of Mathematics and other Essays,* London: Routledge & P. Kegan, 156–198

Paolo Santorio [2012]: "Reference and Monstrosity". *The Philosophical Review*, 121: 359–406

Brian Skyrms [2010]: *Signals*. Oxford: Oxford University Press

Robert Stalnaker [1984]: *Inquiry*. Cambridge (Mass.): MIT Press

— [2002]: "Common ground". *Linguistics and philosophy*, 25(5): 701–721

Ulrich Stegmann (Ed.) [2013]: *Animal communication theory: Information and influence*. Cambridge: Cambridge University Press

Johan van Benthem [2011]: *Logical dynamics of information and interaction*. Cambridge: Cambridge University Press

Clas Weber [2012]: "Eternalism and Propositional Multitasking: in defence of the Operator Argument". *Synthese*, 189(1): 199–219

Timothy Williamson [1994]: *Vagueness*. New York: Routledge

Seth Yalcin [2007]: "Epistemic modals". *Mind*, 116(464): 983–1026