The Philosophical Quarterly Vol. 00, No. 0 *2023*
ISSN 0031-8094 https://doi.org/10.1093/pq/pqad016

HOW TO EXPRESS IMPLICIT ATTITUDES

By Elmar Unnsteinsson[1,2]

I argue that what speakers mean or express can be determined by their implicit or unconscious states, rather than explicit or conscious states. Further, on this basis, I show that the sincerity conditions for utterances can also be fixed by implicit states. This is a surprising result, which goes against common assumptions about speech acts and sincerity. Roughly, I argue that the result is implied by two plausible and independent theories of the metaphysics of speaker meaning and, further, that this is a robust basis on which to make an inference, with a fair degree of confidence, about the relationship between expression and implicit attitudes.

Keywords: speaker meaning, intentionalism, expressionism, implicit attitudes, consciousness, self-deception, speech acts, insincerity.

I. INTRODUCTION

Is what we mean ever determined by our implicit attitudes? If so, how could we ever find out with any degree of confidence? In this paper, I argue that the first question can be answered in the positive. But I do so by employing a specific methodological framework which addresses the second question. There are specific ways in which both the framework and its upshot may be surprising to many readers and, so, I will try to give a brief dress rehearsal before the ceremony.

It is simply common sense, it seems, that if I believe *p explicitly* and I say something to express *p*, my utterance will probably have *p* as its meaning. At least, it will not mean something else, say *q*, even if *q* happens to be one of my implicit or unconscious beliefs at the time of utterance. Similarly, many theorists have argued or assumed that there is a privileged relationship between the contents we express and the contents we explicitly represent in our own minds. Perhaps, however, we are not really asking the right question here. The right question would rather focus on cases where there is some serious potential for conflict or contest between two attitudes in the determination of what we mean. So, if we can plausibly describe cases where the contents of the speaker's

explicit and implicit attitudes are such that either one could, in principle, determine the meaning of an utterance, then we should predict that explicit attitudes are privileged. I will argue that this hunch is seriously undermined by a more detailed consideration of the issues; surprisingly enough, the implicit attitudes are equal or superior contenders in the battle.

As indicated, however, I will not try to establish this by using argumentative strategies that I take to be most familiar to philosophers today. Specifically, I will not argue that the thesis is true intuitively or pre-theoretically, nor that it follows from something most or all philosophers accept already. The thesis is couched in terminology which is too theory-laden for either strategy to work. More specifically, we cannot identify what is common to a set of independently plausible theories and derive interesting conclusions on that basis. Often we mask metaphysical differences between theories by trying to join them together in this manner, using a single expression—like 'meaning' or 'utterance'—which really calls for different interpretations in different frameworks. A different approach is needed.

I present two relatively independent theories from the literature on the metaphysics of meaning. I add to each a fairly standard notion of sincerity and insincerity in speech. And then I try to derive the thesis of interest *from each combination individually*. This yields an abductively robust inference, similar to ones scientists make when they consider many different and incompatible models of climate change to derive as robustly as possible some result of special interest.[1] Now, of course, this methodology has limitations, like any other, and I'll mention one here. I only consider two views on the metaphysics of meaning, and I could have considered more. But because each view needs to be developed in exactly sufficient detail to derive the result, I am limited by considerations of space. However, I happen to think that the two views I have chosen—so-called *expressionism* and *intentionalism*—are more likely than others to carry relatively precise predictions about the underlying mechanisms of linguistic or communicative competence, at least as it bears on the thesis of this paper. Moreover, other prominent frameworks, for example, *conventionalism*, tend to be anti-mentalistic, which alters the dialectic considerably. Such views are sceptical of the role of *any* mental state in the determination of content. Thus, the choice is very far from arbitrary.

In this paper, I will use two 'cases' or 'stories' as part of the overall argument, one about golden bathtubs and another about inebriated elves. The purpose of these cases is not, however, to help the reader consult their pre-theoretic intuitions and reach a philosophically controversial conclusion about the determination of meaning. Instead, the cases are supposed to help us see what follows *from the perspective of the two theories of meaning*. So, the purpose is

[1] Weisberg (2006), Woodward (2006), Dellsén (2017), and Schupbach (2016). The example of climate change is borrowed from Lloyd (2010).

illustrative and theory-laden; we are only consulting our own understanding of the implications of the two theories. Strictly speaking, then, the cases are not parts of arguments for the implication being true, but they are parts of arguments for the implication *being an implication of the theory in question*.

In the next section, I start the discussion by introducing the two theories of meaning as well as a set of interlocking theoretical notions. Here, I must make very specific assumptions about the nature of implicit attitudes which, surely, will not be to everyone's taste. But I hope the assumptions are not too controversial and that those who disagree are happy to accept that I have identified a coherent cognitive category, regardless of whether the labels are exactly right. Even from this perspective, I believe, the results will be interesting. In Sections II and III, I use the difference between *assertive* and *suggestive* speech acts, and the example of buying golden bathtubs, to argue that implicit attitudes may very well determine utterance contents. Next in Section IV, I show how this also applies to the determination of sincerity and compare my example to Freudian slips. Finally, in Section V, I use a different case to argue that the result is not reached simply by engaging our pre-theoretic intuitions.

II. MAKING MEANING

Consider two views about the metaphysics of speaker meaning. These are views about what wholly or partly constitutes the fact that a particular utterance means *p*. On the first view, an utterance means *p* only if it was performed with an intention to produce some *p*-related cognitive effect in a minded creature. As a possible example, by asserting something, I mean *p* only if I intend my act to produce the belief that *p* in my addressee. Call this view intentionalism.[2] On the second view, an utterance means *p* only if the utterance expresses *p*. For an utterance to express *p*, it must stand in some specific relationship to the speaker's own *p*-attitude. One might think, for example, that by asserting something, I mean *p* only if my act is caused by or otherwise indicates my belief that *p*. On some views of this sort, lacking the belief may result in the assertion being a pretend or mock assertion. Call this view expressionism.[3]

Let us make this more precise by introducing some terminology. The two views differ in what they take to be the meaning-maker of particular utterances. The *meaning-maker* involves, at least, a *necessary condition* for an utterance to mean *p* rather than something else. The intentionalist proposes that the meaning-

[2] See, for example, Bach & Harnish (1979), Carston (2002), Grice (1989, 2001), Harris (2022), Neale (1992, 2005, 2016), Schiffer (1972, 1987, 2003), Scott-Phillips (2015), Simons (2017a, 2017b), Sperber & Wilson (1986/1995), Unnsteinsson (2022b), and Wilson & Sperber (2012).

[3] See, for example, Alston (2000), Bar-On (2004), Chomsky (1980), Davis (1992), Devitt (2021), Dummett (1989), Gauker (2003), Green (2007), Hornsby (2000), Horwich (2005), Pagin (2011), Rosenthal (2005a), and Sellars (1969).

maker involves a higher-order state of intention which embeds the *p*-attitude. To simplify, I will focus only on one aspect of this mental state here, the so-called effective or informative intention of the speaker. This is enough to pinpoint the crucial difference between the two views.

Intentionalism

An utterance means *p* only if the meaning-maker involves an intention to produce some attitude *A*(*p*) in the addressee.

The expressionist denies that an intention, or any other pro-attitude, of this sort is a part of a meaning-maker. Instead, the meaning-maker is determined by a first-order state of the speaker.

Expressionism

An utterance means *p* only if the meaning-maker involves some *A*(*p*)-attitude of speaker.[4]

To be clear then, the two views are exclusive, if not exhaustive. The intentionalist rejects expressionism and vice versa. They are formulated as necessary conditions, but they are still just small parts of larger empirical hypotheses about the structure of specific cognitive mechanisms in humans at present, not analyses or partial analyses of any of our pre-theoretic concepts. The 'necessity' in question is thus a conditional one and helps to make the difference between the theories clear. To simplify the discussion, let us assume that there is such a thing as the speech act of assertion and that it is partly defined in terms of its relation to the mental state of belief. Those who disagree may substitute knowledge, or any other state, for belief in what follows. Anyway, the meaning-maker for a *p*-assertion will partly consist, for both intentionalists and expressionists, in a *p*-belief. Going a bit further, I will assume that the meaning of a *p*-assertion has the form: *p*-belief. So, what the asserter means is not only *p*, but the *belief* that *p*. For the expressionist, this is in virtue of the fact that the utterance is caused by or somehow indicates the speaker's *p*-belief. For the intentionalist, it is in virtue of the fact that the utterance stands in the appropriate relation to an intention to produce a *p*-belief in someone, regardless of whether the speaker believes *p*.

For a second piece of terminology, consider the fact that the two views may very well agree about what constitutes the *sincerity-maker* for an utterance. On the intentionalist picture, my *p*-assertion is sincere only if I believe *p*. And the same formulation is available, and plausible, for the expressionist. In the literature on sincerity and lying, this seems as close to a consensus view as possible, so I adopt the view here (e.g. Chan & Kahane 2011; Eriksson 2011; Moran 2005; Ridge 2006; Saul 2012; Schwartz 2020; Stokke 2014,

[4] I follow Harris, Fogal, & Moss (2018) in keeping expressivism in metaethics separate from 'expressionism' as a theory of speech acts.

2018). From this starting point, however, different theorists are lead down very dissimilar paths. The expressionist typically holds that meaning-makers and sincerity-makers are identical, while the intentionalist denies this (e.g. Rosenthal 1989). So, for many expressionists, to assert and mean *p* is to assert and mean *p* sincerely. As mentioned before, it does not follow that the view cannot accommodate insincere assertion; to assert *p* insincerely is to pretend to assert *p* and thereby to pretend to believe *p*. Thus, the speaker does not really assert anything in such a case, although the addressee may very well take the speaker to be doing so. And, of course, the speaker may be producing this appearance knowingly and intentionally. This is the reason, I think, many expressionists argue for what I will call the EXPRESSIONIST COROLLARY. The COROLLARY states that sincerity is more natural or spontaneous than insincerity, in that only the latter, normally, raises specific questions about the speaker's motivation in choosing to speak as they do (Moran 2005: 332; Rosenthal 1989; Sellars 1969; Williams 2002, see also Fricker 1994; Millikan 2017).

It is important to note, however, that on both views sincerity is determined by the attitude-part of what is meant, not the content-part. My (pretend) assertion that *p* is insincere only if my *p*-attitude is of a particular sort, namely one of disbelief. I may have all sorts of other *p*-attitudes compatible with disbelief; the desire that *p* is true, the fear that *p* will be true, and so on. Thus, the sincerity-maker for the utterance depends on the attitude, while the meaning-maker depends on both the attitude and the content. I said that this was important because it is a key factor in the argument I will go on to develop. Roughly, since meaning-makers and sincerity-makers share one crucial ingredient—the relevant mental attitude—we can infer properties of one from properties of the other.

Now, the third and final piece of technical terminology is the notion of *activation preference*. Someone's mental state is *activated* if and only if it is ready to exert causal influence on their behaviour and cognition (see, e.g. Frankish 2004: 14–7). My belief that green means go is thus activated when I drive my car. Plausibly enough, mental states can vary in their degree of activation; my desire to drink Guinness can consume my whole existence, making it impossible to think about anything else; or it can be a mere understudy, ready to take the stage in the absence of stronger desires. But there are also qualitative differences in activation. Some belief is activated *implicitly*, let us say, if it is ready to exert causal influence while the speaker is *not* disposed to self-ascribe the belief. Arguably, there are many examples of this kind of activation from theories of implicit bias and self-deception (Funkhouser 2005; Funkhouser and Barrett 2016; Gendler 2008; Quilty-Dunn & Mandelbaum 2018; Schwitzgebel 2010; Unnsteinsson 2022b: ch. 3; Wilson 2002). To illustrate, if I am self-deceived about being bald, then I may very well believe, deep-down, that I am really bald, but I would not say that I am, to myself privately or to others, at least not sincerely. The activation of an actual belief is *explicit*

just in case it exerts causal influence in the normal way, *and* I am disposed sincerely to self-ascribe the belief. So, for the purposes of this paper, activation is of three qualitatively different kinds, the state is (1) only implicit, (2) only explicit, or (3) both. Note that this allows for the possibility of beliefs or belief-like states that are merely explicit as in category (2), that is to say, states that consist in nothing over and above the relevant disposition to self-ascribe that very state.

This qualitative notion of activation is well supported by empirical work on bias and implicit cognition, although it is not uncontroversial. Here I use the unconscious production of misleading evidence for illustration, because it marks the closest empirically-attested relative to the cases I will present later (Funkhouser and Barrett 2016: 688–9). Humans commonly modulate the pitch of their voice to further their social goals, seemingly without awareness. People do this both when speaking to someone to which they are attracted and when they offer their expert advice to someone. Let's focus on the latter. Professionals will lower their voice pitch when speaking in the role of experts, as compared to the more mundane context of being asked for directions (Sorokowski *et al.* 2019). It is plausible to construe this as the unconscious production of misleading evidence about the speaker's normal voice pitch. The modulation seems socially motivated—it's a form of impression management—because peoples' attitudes and reactions vary in response to differences in pitch. A lower voice signals expertise and authority. Admittedly, this example allows for a range of different interpretations but, still, there must be *some* attitude which guides speech production, and is not explicitly self-ascribed. Perhaps it is the belief or belief-like attitude that *this* is the speaker's normal voice pitch. This qualifies as implicit or unconscious in activation as these notions will be understood here.

The notion of activation *preference* is the idea that some phenomenon of interest may prefer one type of activation over another. That is to say, in cases of conflict or contest between the two types of activation, the phenomenon of interest may be constitutively determined by one as opposed to the other. And this may be a completely general truth about the phenomenon, namely, the preference might be a stable and robust one. To take a classic example, consider the intuitive truth conditions of attitude ascriptions as a phenomenon of interest. It seems false to say (i) that Lois Lane believes that Clark Kent can fly. But still, it seems true to say (ii) that she believes, of the person we call Clark Kent but she calls Superman, that that person can fly. So, perhaps, the intuitive truth condition of (i) prefers to be determined by explicit activation of the relevant belief in Lois' mind. That is to say, Lois is not disposed to self-ascribe any belief with a sentence like 'Clark can fly'. The truth of ascription (ii) implies that the belief in question is at least implicitly activated, but this is not sufficient for the intuitive truth of (i). Or so one might argue (e.g. Unnsteinsson 2022b: ch. 1).

III. IMPLICIT MEANINGS AND GOLDEN BATHTUBS

Here, instead, I will argue that the meaning-maker of an utterance prefers one type of activation. It follows, as I will show in the next section, that the sincerity-maker does so too. But first we need a better understanding of the notion of conflict or contest between types of activation. In the abstract, the idea is that a particular action can be motivated or explained in terms of two attitudes at the same time, such that the attitudes seem somehow inconsistent, where one is activated implicitly and the other explicitly. This is best explained with an example. Let us assume that there is a distinction between *assertives* and *suggestives*, roughly along the lines proposed in Bach and Harnish (1979: 42–3). I simplify the distinction and leave out unnecessary detail, partly so it can be accommodated both by expressionists and intentionalists.

Assertive

S means *p* assertively, by uttering *X*, only if what *S* means is a belief that *p*.

Suggestive

S means *p* suggestively, by uttering *X*, only if what *S* means is (a belief) that there is a reason to believe that *p*, but that the reason is insufficient.

So, for example, if *X* is the English sentence

(1) A golden bathtub would be nice,

then the speaker can, depending on the context, mean *X* assertively or suggestively. Imagine that we are trying to figure out what to give to our friend for her birthday. In some contexts of that sort, (1) could be meant assertively, and in others, merely suggestively. The latter utterance might be a part of listing things, just to have some options to consider. The former would simply be the speaker's way of expressing or communicating the belief that our friend ought to have a golden bathtub, and nothing else.

Plausibly, there is a similar distinction between what Bach and Harnish (1979: 47–8) would call *requirements* and *advisories*.

(2) Buy her a golden bathtub.

In some contexts, I might utter (2) only to express or communicate that it is advisable to buy the bathtub, and in others, what I mean might involve a strict command or order.

Both expressionists and intentionalists can accommodate these examples, in slightly different ways. But that point need not detain us here. What I want to note about the examples is that a speaker can be self-deceived about whether they mean something assertively or suggestively, or whether they mean it as a requirement or as an advice. And now we have the tools to describe this type of self-deception in some detail. Consider a context for (1). I firmly believe that our friend, Peg, needs and should have a golden bathtub. To my mind,

it is a no-brainer, and I am not inclined to take any other suggestion very seriously. But golden bathtubs are expensive, frivolous emblems of aristocratic thoughtlessness. And so, I would never deliberately and sincerely self-ascribe the belief that someone needs or ought to have a golden bathtub. I would feel ashamed if I would, maybe in a moment of carelessness, perform a speech act whereby I represent myself as believing such a thing. This is certainly paradoxical or close to inconsistent; I did utter (1)! But I self-deceptively think of (1) as a suggestion and not an assertion. And what I lack is the disposition to self-ascribe the relevant belief.

This kind of example fits nicely within the category of deceptive behaviour already discussed, namely the production of misleading evidence. Here I do not produce misleading evidence about my normal voice pitch, but about my actual attitude in producing the speech act. Going along with Funkhouser & Barrett's (2016) description of this category, the evidence can be unconsciously aimed at misleading myself and others. If this is right, then it is a form of highly flexible and strategic deception, even if it is unconscious, because it is sensitive to changes in the speaker's assumptions about the context and addressee (see Doody 2017 for a contrasting view). More specifically, it is plausible that the sentence (1) is selected partly because it does not clearly distinguish between assertive and suggestive speech acts. This would be explained by the postulated implicit or unconscious motivation to express belief in *p* and not anything weaker. I unconsciously mislead myself, and others, into thinking that I do not believe *p* or do not intend to produce that belief in the addressee.

This seems to court controversy on two fronts. First, I appear to assume that implicit attitudes must have propositional structure (Mandelbaum 2016 would agree, Madva 2016 disagree). Secondly, deflationary theories of self-deception—where it is reduced to motivated belief (e.g. Mele 2001)—seem to be ruled out. But these assumptions are not strictly necessary. Roughly, we could have a view which restricts the *belief that p* to a mental state which is in no way determined by the speaker's acts of self-representation as a *p-believer*. And so, the apparent inconsistency in self-deception would really be between different kinds of things; beliefs and (dispositions to perform) acts of self-ascribing beliefs, and deflationism remains possible (see, e.g. Bach 1981). Moreover, even if this is true, we could still argue about whether it is necessary to think of the underlying belief-state as having propositional structure. It is consistent with this view, at least, that assigning such a structure is an idealisation which we may want to discard as our knowledge advances.

Here it becomes helpful to distinguish between the intentionalist and the expressionist, because self-deceptive speech makes for interesting differences. Start with the intentionalist. On this picture, I take myself to have uttered (1) suggestively, thinking that I thereby intend to produce in you the belief *that there is a reason, albeit insufficient, to believe p*. This is not a belief I myself have, because what I really believe is *p*, full stop. But you are my personal assistant

and you know that it is your job to buy the present for Peg. And, of course, I already know that this is your job. In this case, it seems like you understand me correctly only if you take me to be expressing my belief that *p*, even if I self-deceptively think of myself as making a mere suggestion. As my personal assistant, you will have gotten it wrong if you buy a gold chaise lounge instead. But this should have been perfectly consistent with a mere suggestion.

Admittedly, the details are extremely delicate. I am not arguing here that there are no possible variations on this kind of example, which would appear the same on the surface, but with significant differences in the mental state of the speaker. For example, I might not suffer from any self-deception about my beliefs or intentions in uttering (1). In that case, I might intentionally express a suggestion and thereby intend to implicate something stronger. Perhaps my social embarrassment is obvious and known to all, and pretending to make a mere suggestion might be part of my effort to save face. But it is equally true that speakers could be self-deceived about these kinds of things, and that is the kind of example I want to focus on here. In that kind of example, it seems like the personal assistant may understand my utterance by coming to recognise my *implicit* belief that *p*. It is implicit in the sense that I am not disposed to represent myself sincerely and explicitly as a *p*-believer, to anyone or to myself. It follows that the meaning-maker of my utterance, on the intentionalist view, has an implicit activation preference.

The expressionist will have to say something very different. On her view, assertive utterances of (1) are truthfully prefixed by 'I (firmly) believe that...' and suggestive utterances by 'I (merely) speculate that...'. Now, we are assuming a certain activation profile for my belief that a golden bathtub would be nice (for Peg). I hold this belief implicitly but not explicitly, as those terms were defined before. So, almost by definition, if I utter (1) suggestively—without the appropriate prefix—then I misrepresent my own beliefs. To that extent I am deceived about my own doxastic state. That is to say, I am disposed to represent myself as only speculating that *p*, while in truth I believe that *p*. Let us represent those two mental states as $B(p)$, for the belief, and $S(p)$, for the state of speculation. And, importantly, we assume that a particular utterance of (1) is perfectly suitable for the direct expression of either state, depending on the actual doxastic profile of the speaker.

Now, even if (1) is suited for the expression of $S(p)$ we are also assuming that it is false that I speculate that *p*. It follows, on the expressionist view, that I cannot sincerely mean $S(p)$ by uttering (1), since I can only express my actual mental states sincerely. Is it possible that I am *pretending* to mean $S(p)$ by uttering (1)? Sure, and this would be consonant with Tamar Gendler's (2007) idea that self-deception should be understood as a form of pretense more broadly. But, this strategy gives rise to a serious objection. The point of the objection is not to show that pretense theories of self-deception are false, only that, for the kind of case we have constructed, it is unavoidable to conclude

that the speaker in fact expresses their implicitly held belief, rather than the explicitly represented speculation. Pretense theorists could consistently accept this conclusion without giving up the essentials of their theory.

The objection is as follows. Assume that some utterance *X* is equally compatible with the speaker directly meaning *B*(*p*) and *S*(*p*) in the sense that the speaker could have directly meant either one while holding everything else about the context fixed. Then, if expressionism is true and the speaker is trying to be sincere, then the speaker will express the *p*-attitude, which happens to be their actual *p*-attitude. By assumption, this will be *B*(*p*) and not *S*(*p*). The implicit belief slips out, again, somehow unnoticed by the self-deceived speaker. How could this possibly be true? Well, think about what sincerity consists in for the expressionist. The speaker makes an utterance to express their attitude to *p*. Since they are trying to be sincere—or, better, they have not decided to deviate from the normal sincere response—they will want to express their actual *p*-attitude and not any *p*-attitude that they do not have. So, it follows in this case that the speaker does *not* want to express *S*(*p*), but they in fact do want to express *B*(*p*). This should suffice to give the latter state the causal and explanatory advantage, that is to say, in cases of conflict, it is more plausible to think that the speaker expresses a state they want to express rather than a state they do not want to express, when they are trying to be sincere.

Going along with pretense theory, we can conclude that the speaker's pretend-meaning is the speculation that *p*, but that the actual meaning is the belief that *p*. But the expressionist is also free to drop pretense theory and argue that the speaker does not pretend to mean anything and that the speaker's self-deception about what they mean has a different explanation. They might be disposed, for example, to tell themselves explicitly that what they meant was only *S*(*p*) and not *B*(*p*); 'I hope she doesn't take it seriously, it was really just a suggestion.' It is worth emphasising, yet again, that this makes the examples under discussion highly specific. The argument would not work for cases where the speaker believes not-*p* and expresses and means the belief that *p*. It is possible to utter 'It's raining' to mean directly your belief that it is raining—and maybe also your disappointment or annoyance—but, normally, you cannot utter that sentence to directly express and mean your *disbelief* that it is raining. This remains true even if the hearer may detect disbelief in your tone of voice, for example. And we can leave indirect speech acts to one side here, because the focus is on what speakers mean directly or literally, in a way that is intuitively compatible with the encoded meaning of the expression uttered.

Now we can finally understand what the contest or conflict of attitudes consists in, on both expressionist and intentionalist views. Meaning-makers for utterances may give rise to contests between implicit and explicit attitudes with the same contents, when the utterance is strongly compatible with both attitudes. In fact, however, we have seen that the contest seems to favour

the implicit attitude over any conflicting attitude which is explicit. We should pause to note that this ought to be surprising or even a bit paradoxical. How do self-deceived speakers then manage to express and mean what they self-deceivingly believe? If the implicit attitude can somehow slip under the radar and burst open in normal conversation, how can it be implicit at all? Well, as I have already stated, the contest only happens in special situations. When there is no contest, the question of activation preference does not arise, and speakers can easily express and mean whatever they merely believe explicitly. And this is well and good, because it is at least possible that some belief-like states besides the self-deceptive ones—e.g. beliefs that are essentially encoded in some natural language sentence, or beliefs with no other effect on behaviour or cognition—consist in dispositions to self-ascribe that belief explicitly.

IV. SINCERITY, CONSCIOUSNESS, AND FREUDIAN SLIPS

We now have enough wind in our sails to reach the desired conclusion about the connection between meaning-makers and sincerity-makers. One of my objectives was to establish the claim that if meaning-makers prefer to be implicit, then sincerity-makers do so as well. And this connection is supposed to hold for both expressionists and intentionalists. On this point, we will start with the former, because the connection is so obvious. As already indicated, the expressionist holds that the meaning-maker of an utterance is identical to its sincerity-maker. The utterance *u* expresses the belief that *p* only if *u* stands in the right relation to the speaker's belief that *p*. Insincere expression is a matter of pretense. On one view of this kind, the belief must cause the utterance in the right way. Since they are identical, if the meaning-maker prefers to be implicit, then it follows that the sincerity-maker does too.

Intentionalism only adds one small wrinkle. Meaning-makers are not identical to sincerity-makers but, rather, they are partly constituted by higher-order intentions which embed the sincerity-relevant attitude. We have already established that one part of the meaning-maker, for the intentionalist, prefers implicit activation. More precisely, this is the attitude-part of the *A*(*p*)-attitude which is embedded by the speaker's so-called effective intention. In uttering (1), the speaker will express and mean the implicitly held *B*(*p*)-state in preference to the explicitly held *S*(*p*)-state. The intentionalist agrees with the expressionist in assuming that the speaker is insincere in meaning *A*(*p*) by utterance *u*, only if the speaker does not bear the *A*-attitude to *p* in making *u*. It follows, then, that the sincerity-maker for *u* is partly constituted by the speaker's *A*(*p*)-state. If so, then, even for the intentionalist, it is true that if (one specific part of) the meaning-maker prefers to be implicit, then the sincerity-maker prefers to be implicit as well.

Again, I should emphasise the specificity of the case at issue. For the intentionalist, there are additional ways in which to make the example into a non-contest. Possibly, the speaker *intends* to communicate the *S*(*p*)-attitude and intends *not to* communicate any other *p*-attitude, even in a case where uttering (1) is also compatible with an intention to mean the *B*(*p*)-attitude. If this is possible, there will be no contest or conflict, regardless of the speaker's implicitly held belief that *p*. The case I have described is slightly different. Because the example is one where I am self-deceived about my attitude to *p*, such that I am disposed to self-ascribe *S*(*p*) while, deep down, I know that my state is a *B*(*p*)-state, it is plausible to think that I have a hidden intention to communicate the latter. I am holding back because of perceived social pressure and fear of embarrassment. In reality, however, I am the kind of person who thinks that golden bathtubs are nice presents.

Intentionalists are allowed their own moment of studied perplexity. Many theorists in that tradition have argued or assumed that what speakers express and mean directly is going to be identical to what they consciously or explicitly believe, at least if such a state is in the offing. Normally, however, this view is taken to follow from a more general principle of privileged access; normal speakers have immediate access to what they intend and mean, while addressees must infer the meaning on the basis of the utterance (e.g. Fodor and Lepore 2004: 84; Neale 2005: 179–80; Schiffer 1992: 515; 2016: 498–501).[5] And, certainly, this seems more or less plausible. But only more or less. If it were a robust generalisation we ought to expect the opposite result from what we have seen so far. That is to say, we ought to predict that the explicitly represented and intended *S*(*p*)-attitude would easily win a competition with the implicitly held *B*(*p*)-attitude. Of course, as already mentioned, it is crucial for the intentionalist contest that both attitudes are, in some sense, intended by the speaker. The surprise comes from the realisation that the implicit attitude and hidden intention are even in the running.

Admittedly, I have argued that the addressee will understand the utterance correctly only if they identify the implicit attitude, but this kind of understanding comes in degrees, especially for the intentionalist. In this sense, the addressee understands the utterance better if they recognise that I self-deceptively intend one thing while I really intend another. That is, the addressee then understands the utterance-act *better* in the sense of having a fuller explanation of why and how it was produced, by grasping the mental state of the agent in more detail. Since I did not make as if I were merely suggesting something to thereby implicate something stronger, the addressee will misunderstand if

[5] It should be noted that there are very many examples of authors arguing or assuming that the *interpretation* of speech acts is implicit or unconscious (e.g. Bach & Harnish 1979: 93), but here I am only concerned with the idea that the *speaker's* own attitude—which is supposed to be partly constitutive of their meaning—can be either implicit or explicit.

that is what they think I was doing. So, there is a real, but very fine, distinction between meaning something by implicature and meaning something without taking oneself to be doing so. Note, however, that this is due to a special feature of the example, namely, that the utterance of (1) is evidentially compatible with the speaker directly meaning one or the other, and does not require one of the attitudes meant to be meant indirectly. Even so, it should be a surprise that the implicit attitude seems relevant to the interpreter's success.

At this point it might seem that the conclusion of my argument could have been reached without so much effort. After all, there are plenty of theorists who object to the idea of privileged access (e.g. Carruthers 2011; Schwitzgebel 2006; Unnsteinsson 2022a). Moreover, if there is such a thing as a Freudian slip—where I reveal my hidden intention to myself and others—it would seem easy to find relevant counterexamples. Let us take these suggestions in turn. The thesis I have been arguing against is, simply, that what one means or expresses is always conscious or explicit. One need not endorse the view that we have privileged access to our own mental states to be sympathetic to that idea, because it seems so commonsensical. Indeed, this appears to be the dialectical position of many expressionists, David Rosenthal being the most obvious example. He argues against introspection as a privileged form of access (2005b), but also claims that verbally expressed thoughts must be conscious (1998, 2005c).[6]

I will not object to Rosenthal's argument here, as it directly depends on his theory of consciousness and its connection to verbal expression. Roughly, he believes that the best explanation for why speakers can spontaneously tell others about their own mental states must appeal to higher-order thoughts about those very states. Since having such higher-order thoughts about one's own thought is constitutive of consciousness, on his view, verbal expression is sufficient for consciousness. Significantly, however, there are exceptions, even on Rosenthal's picture. Verbal expression is not sufficient for the consciousness of *affective* or *higher-order* states, but he argues that his theory provides satisfying explanations in both cases (1998, 2005c). Now I can explain how this is relevant to my own argument. This connection between consciousness and expression is taken, by Rosenthal and others, to be common sense (Rosenthal 2005c: 282, n. 2). On this basis, it is supposed to speak in favour of the proposed theory that it explains the common sense generalisation, while accounting for the exceptions as well.

My argument is importantly different. Now, if indeed expression and sincerity privilege explicit or conscious attitudes, we would predict that explicit

[6] Other expressionists have endorsed related views. For example, Mitch Green holds that what is expressed '... must be of a sort that can be known introspectively' (2007: 39). Of course, many philosophers have argued for 'first-person authority' of this kind without being as easily categorised as expressionists or intentionalists (see, e.g. Davidson 1984; Heal 2002).

attitudes are preferred in cases of contest, as this notion has been understood here. But, on the contrary, we get the rather surprising result that, in cases of contest, implicit attitudes are at least equal contenders, if not superior. Even the former, weaker conclusion would be sufficient. That is to say, the common sense prediction would not be that implicit attitudes are equal contenders in cases of contest, because that would not be consistent with the idea that conscious attitudes are *privileged*. I think this conclusion warrants a higher credence than we would otherwise have assigned to the claim that sincerity and utterance content are both determined by implicit attitudes, rather than by explicit attitudes, *whenever* the former are available and explanatorily relevant.

To see this point, I think, it helps to consider Freudian slips briefly. Genuine Freudian slips have a set of highly specific features, which makes generalisation less robust. Most actual cases are best explained—at least in part—in terms of phonetic and phonological factors. For example, the so-called lexical bias effect predicts that speakers are much more likely to perform speech errors where a familiar word replaces the target expression. So, I'm already more likely to utter 'Osama' when making a mistake in trying to utter 'Obama', than any other similarly sounding sequence. In such cases, it is easy to assign the content actually intended by the speaker, as if they had made a simple error in pronunciation (Unnsteinsson 2017).[7] Moreover, speech production errors of this kind normally involve subsentential expressions, rather than full sentences. And so, the Freudian explanation, apart from postulating mechanisms of repression, comes awfully close to a claim about what speakers may reveal about themselves by their choice of words. And, of course, we reveal all sorts of information without that information becoming the content of what we mean or express. Anyway, this is not to deny the possibility of a genuine Freudian slip. But when a Freudian slip is motivated by a hidden intention to express some proposition—rather than by some hidden association of two phrases—'golden bathtub' and 'golden shower', say—the intention will tend to become explicit, without competing with a different intention in the determination of content. In a moment of clarity, I might 'slip' and say to my friend, 'I don't like you', only to realise that following up immediately with 'I didn't mean to say that' is not very convincing. The same point can be made in terms which are more appealing to the expressionist.

The bottom line, however, is that Freudian slips are not a very powerful reason to doubt the common sense view that, normally, we only express our explicit attitudes. As we have seen, the common sense view could potentially survive apparent counterexamples of this kind by providing case-based and more specific explanations. Sure, I have also constructed highly specific circumstances, involving a contest between attitudes, but this does not undermine the argument. My examples do not involve performance error and seem not

[7] See Reimer (2004) and Stokke (2014) for a contrasting view.

to appeal essentially to specific types of mental states (e.g. affective or higher-order ones). And so, as I will illustrate in the next section, the argument seems to be a good basis on which to generalise.

V. INTUITION AND GENERALIZATION

How can I be so confident that the implicit attitude is in the running in every contest of this general kind? Have I done anything more than appealed to my own intuition that what the speaker means and expresses by uttering (1) is the hidden belief rather than the explicitly represented state of speculation? Others will surely have the opposite intuition in this particular case.

In this section and the next, I want to guard against this reaction. I have tried to show that expressionists should, by the light of their own theory, predict that the implicit attitude rather than the explicit attitude is expressed, in the kind of case at hand. This is worth spelling out as clearly as possible. The speaker is assumed to satisfy whatever the theory requires to guarantee that the utterance is sincere. For the expressionist, to assert *p* sincerely is (at least) to assert *p* and believe *p*. Expressionists also tend to hold that insincere speech requires deliberate deviation from the most immediate or natural course of action (the so-called EXPRESSIONIST COROLLARY). That is to say, you withhold what you believe to be true, deliberately and intentionally. In our example, I utter (1) to express my actual *p*-attitude. My *p*-attitude happens to consist in *the belief that p*. I self-deceptively represent this attitude as one of mere speculation about the truth of *p*. Now, I do not have any deliberate intention or desire to express or communicate an attitude that does not match my own attitude. It follows, since the utterance is compatible with both the direct expression of belief and the direct expression of suggestion or speculation, that what I express is my belief, and not my speculation.

Here it will help to introduce a new kind of example. Let us assume that Gunnvör believes in elves. She treats elf-rocks with care and respect and finds herself in agreement with efforts to keep the disruption of alleged dwellings of elves to a minimum. So, she is one of the roughly third of the population of Iceland which, it is alleged, have this belief. But still, she does not want to appear superstitious and tends not to think of herself, explicitly at least, as a believer. If the matter comes up, privately or in the company of others, she is much more disposed to scoff and take offence at the suggestion that she believes in elves. But Gunnvör really wants to dance at The Tipsy Elf, where belief is a requirement for entry. So, intending to lie, she says to the gatekeeper,

(3) I believe in elves.

And Gunnvör can dance the night away. Did she lie? Was she insincere? Maybe the answer is unclear or even indeterminate. Interestingly, the expressionist seems to have a good explanation. By assumption, Gunnvör satisfies

one condition for sincerity, but does not satisfy its COROLLARY. She definitely expresses a belief that she in fact has. But she was trying to be insincere and, to the extent that trying makes it so, she is (a bit) insincere. Perhaps the COROLLARY—insincerity as deliberate deviation—is optional or merely symptomatic. If so, we will have to say that Gunnvör was unintentionally sincere. If that is right, then the meaning-maker for (3) must be her implicit belief and, it will follow, the utterance is sincere in virtue of a match between an implicit attitude and the attitude expressed. Notice, for example, that this would help to explain the fact that Gunnvör herself might come to realise, after the utterance, that she really does believe in elves. She might have some distinctive phenomenology associated with telling the truth after a long time of keeping silent or being evasive.

The other part of my argument was to say that we can arrive at the same conclusion from intentionalist premises. The intentionalist makes no assumption that sincerity is natural and insincerity is a deliberate deviation, because the meaning-maker is an intention to produce an attitude in the addressee, regardless of how it matches up with the speaker's own attitudes. Still, insincerity is determined by a mismatch and, I argue, the mismatch prefers the implicitly activated state to the explicit one. This can now be illustrated with our new example, although the description will involve some interesting complications.

According to the intentionalist, Gunnvör utters (3) intending thereby to produce, in the mind of the gatekeeper, the belief that she (Gunnvör) believes in elves. She takes herself to be lying, because she takes herself neither to believe in elves, nor to believe that she believes in elves. By assumption, however, she does implicitly believe in the existence of elves, but she self-deceptively expresses and represents herself otherwise. So, if the sincerity-maker is the belief that elves exist, it may seem like Gunnvör is sincere in virtue of an accidental match between what she intends to communicate and what she believes implicitly.

This is too quick, however. As I have described the example we seem to lack a match between (i) what the speaker means and asserts and (ii) what she implicitly believes. The attitude Gunnvör intends to produce, and thus constitutes what she means according to this theory, is a higher-order belief, namely (4).

(4) the belief that Gunnvör believes in elves.

And (4) is not identical to the hypothesised implicit belief, namely (5).

(5) the belief that elves exist.

Arguably, the problem remains for the intentionalist even if we switch to a more basic utterance, for example (6).

(6) Elves exist.

According to one strand in intentionalist theory, which we can call 'exhibitionism', the attitude Gunnvör would most likely intend to produce by a literal

utterance of (6) in this particular context will still be the higher-order belief in (4).[8] On the exhibitionist construal, (6) is a direct expression of (4) rather than (5). This is because it is already common ground in the context that the gatekeeper believes in elves and, further, that the purpose of Gunnvör's utterance of either (3) or (6), in the context, is primarily to persuade the addressee that Gunnvör shares this belief. Gunnvör intends the utterance primarily as an exhibition of her own belief, as there is no need to convince the gatekeeper that elves exist.

The problem here is that it seems very implausible that (4) captures the content of Gunnvör's *implicit* belief and (5) is a much better candidate. Some would even argue that (5) captures the implicit belief while the *absence* of (4) in her mind is part of what constitutes her self-deception (see, e.g. Funkhouser 2005). Now, I don't need to argue with any of this. It is perfectly possible that some states of self-deception consist in this type of mismatch between first-order and higher-order beliefs. But this is not necessary. As I have described the example, Gunnvör simply believes that elves exist but has a disposition to represent herself, to herself and others, as not believing so. And she takes herself to be sincere in so representing herself. As far as I can see, the higher-order belief is explanatorily optional and so I remain agnostic about whether her disposition constitutes or otherwise requires a higher-order belief of the relevant sort.

But we are not out of the woods yet. Exhibitionism rules out the simple option of saying that Gunnvör is sincere in virtue of an unintentional match between her implicit attitude and the attitude expressed or meant. This is because we are assuming that (4) is roughly what she means by her utterance but (5) is the content of her implicit belief. There are some options open to exhibitionism here. For example, one could say that Gunnvör implicitly believes *both* (4) and (5).

A better strategy is to reject exhibitionism. Even if it is true that (4) captures one effect Gunnvör definitely intends to produce in the mind the of the gatekeeper, this is parasitic on the production of some other effect. And that other effect is something very much like (5). To theorise in terms of nothing but belief is a massive idealisation which becomes less useful when we look under the hood and get our hands dirty. What exhibitionism gets right is that the speaker, in this kind of case, does not intend simply to produce a *p*-belief in the addressee but, rather, something slightly different. More specifically, Gunnvör intends to *remind* the gatekeeper or *call their attention to* the existence of elves (6), or her belief in elves (3). She does this, certainly, in order to persuade the gatekeeper that she believes in elves, regardless of whether she utters (6) or (3), in the context as described. The important point, however, is that this adds

[8] See Grice (1969: 106–12). For discussion, see McDowell (1980: sec.5), Neale (1992: 545–7), and Wharton (2009: 25).

a few mains to our menu. If I remind you that it is 3 p.m., I am definitely insincere if I do not believe that it is 3 p.m. This remains so even if you already know the time, and I believe that you know the time, and so on.

Obviously, this can get complicated very quickly, but the details need not detain us. To generalise, the alternative to exhibitionism is to say that the speaker intends to produce some *p*-attitude in the addressee, by making an utterance, and they are insincere only if they lack some appropriately related *p*-attitude. So, I may *remind* you that *p* while lacking the *belief that p* and this may be sufficient to determine the insincerity of my reminder. And this remains true even if the speech act of reminding has additional, special features, making it distinct from asserting. Both will still have beliefs as their sincerity-makers or parts thereof. Finally, if this analysis is applied to the case at hand, we can say that Gunnvör is unintentionally sincere in virtue of her intention to call the gatekeeper's attention to her belief in elves, thinking of herself as having no such belief while, in fact, she does believe in elves.[9]

In this section, I have argued that two different theories of the metaphysics of meaning lead to the same conclusion about the activation preference of sincerity-makers. Both theories predict that, under conditions of contest between two candidate attitudes, the implicit attitude is preferred in fixing the sincerity conditions of the utterance. I believe that this is a surprising prediction, because many philosophers have argued or assumed that insincerity consists in a mismatch between what is expressed or communicated and the speaker's *explicit* attitudes (e.g. Stokke 2014, 2018). This is evidence for the more general conclusion that what speakers mean and express is determined by their implicit attitudes in preference to their merely explicit ones.

As we have seen, there are other views about the metaphysics of speaker meaning than the two considered here. But the choice is a principled one. Expressionism and intentionalism are the only prominent views which are purely *mentalistic*, while others are anti-mentalistic or some hybrid. Purely mentalistic views take the content of an utterance on an occasion to be determined directly by the speaker's mental state at the time of utterance. Others will reserve a role for notions like convention, context, or function in the determination of utterance content. Certainly, mentalistic frameworks may also need such notions, but not when answering the question of what strictly constitutes speaker meaning. It does not follow that anti-mentalistic views cannot distinguish between implicit and explicit meaning, only that the underlying explanation will be structurally different. Take a simple form of conventionalism for example. On such a theory, perhaps, a contest between assertion and suggestion may consist in the fact that the speaker somehow participates in different ways in two distinct conventions at the same time. Or it is indeterminate which convention is operative in the context. A more pessimistic view would be that the

[9] Jessica Keiser (2022) emphasises this aspect of a Gricean communicative intention, namely that the speaker directs the hearer's attention to some content or meaning.

difference is not part of speaker meaning at all.[10] So, the point here is only that the argument developed in this paper would have to look very different if such a view is taken as the starting point. I suspect, however, that if the argument works in a mentalistic framework, it can be extended to some of the less mentalistic frameworks. But, I cannot establish this here.

VI. CONCLUSION

A simple way to understand the argument in this paper is to think of it as an elaborate booby trap. In fact, it's a trap within a trap. First, we need utterances that speakers can produce to express two different attitudes to the same proposition. Second, we need a speaker who is conflicted, such that one attitude is implicit and another is explicit. This is the raw material for the first trap. In making the utterance, the speaker may accidentally reveal their hidden beliefs to the world. Nothing much hangs on the notion of accident here; it only means that, in relative terms, the speaker did not deliberately plan for this to happen.

The second trap is for the theorist. Many theorists assume that explicit (or 'conscious') attitudes are privileged in the determination of speech act content. But if this were true, we would predict that in any conflict of this kind, the explicit attitude would be the clear winner. On reflection, however, this seems not to be true. More importantly, if we consider two prominent theories of speech acts, it seems that both would predict the opposite conclusion. Finally, when we consider the consequences that this has for adjacent phenomena—here I have focused on sincerity conditions—we obtain results that may seem a bit counterintuitive. Roughly, if sincerity were determined by our explicit beliefs, we should expect the conflict to privilege those. And many philosophers believe sincerity is a matter of conscious or explicit attitudes. But they will step right into my trap.[11]

REFERENCES

Alston, W. P. (2000) *Illocutionary Acts and Sentence Meaning*. Ithaca: Cornell University Press.
Austin, J. L. (1975) *How to Do Things with Words*. Oxford: Clarendon Press.
Bach, K. (1981) 'An Analysis of Self-Deception', *Philosophy and Phenomenological Research*, 41: 351–70.

[10] These are speculations on behalf of a variety of conventionalists. The type of framework I have in mind belongs to a very broad church. See, for example, Austin (1975), Lepore & Stone (2015), Reimer (1992), Searle (1969), and Stojnic (2021).

[11] Many thanks to David Plunkett and Hrafn Ásgeirsson for comments on an earlier version. Thanks also to the participants in the 2022 Iceland Reference Workshop and the University of Reading Philosophy Colloquium, in January 2023, for their questions and comments. Finally, I wish to thank two anonymous reviewers for this journal for very helpful comments and suggestions. This work was supported by the Icelandic Research Fund, grant no. 206551.

Bach, K. and Harnish, R. (1979) *Linguistic Communication and Speech Acts*. Cambridge: MIT Press.
Bar-On, D *et al.* (2004) *Speaking My Mind: Expression and Self-Knowledge*. Oxford: OUP.
Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: OUP.
Carston, R. (2002) *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Oxford: Blackwell.
Chan, T. and Kahane, G. (2011) 'The Trouble with Being Sincere', *Canadian Journal of Philosophy*, 41/2: 215–34.
Chomsky, N. (1980) *Rules and Representations*. New York: Columbia University Press.
Davidson, D. (1984) 'First Person Authority', *Dialectica*, 38/2–3: 101–12. https://doi.org/10.1111/j.1746-8361.1984.tb01238.x.
Davis, W. (1992) 'Speaker Meaning', *Linguistics and Philosophy*, 15/3: 223–53.
Dellsén, F. (2017) 'Abductively Robust Inference', *Analysis*, 77/1: 20–29. https://doi.org/10.1093/analys/anx049.
Devitt, M. (2021) *Overlooking Conventions: The Trouble with Linguistic Pragmatism*. Frankfurt: Springer.
Doody, P. (2017) 'Is There Evidence of Robust, Unconscious Self-Deception? A Reply to Funkhouser and Barrett', *Philosophical Psychology*, 30/5: 657–76.
Dummett, M. (1989) 'Language and Communication', in A. George (ed.) *Reflections on Chomsky*, 192–212. Oxford: Blackwell.
Eriksson, J. (2011) 'Straight Talk: Conceptions of Sincerity in Speech', *Philosophical Studies*, 153/2: 213–34. https://doi.org/10.1007/s11098-009-9487-2.
Fodor, J. and Lepore, E. (2004) 'Out of context', in *Proceedings and Addresses of the American Philosophical Association*, 78: 77–94.
Frankish, K. (2004) *Mind and Supermind*. Cambridge: CUP.
Fricker, E. (1994) 'Against Gullibility', in A. Chakrabarti and B. K. Matilal (eds.) *Knowing from Words*, 125–61. Dordrecht: Kluwer Academic Publishers.
Funkhouser, E. (2005) 'Do the Self-Deceived Get What They Want?', *Pacific Philosophical Quarterly*, 86/3: 295–312.
Funkhouser, E. and Barrett, D. (2016) 'Robust, Unconscious Self-Deception: Strategic and Flexible', *Philosophical Psychology*, 29/5: 1–15.
Gauker, C. (2003) *Words without Meaning*. Cambridge, MA: MIT Press.
Gendler, T. S. (2007) 'Self-Deception as Pretense', *Philosophical Perspectives*, 21/1: 231–58. https://doi.org/10.1111/j.1520-8583.2007.00127.x .
——. (2008) 'Alief and Belief', *The Journal of Philosophy*, 105/10: 634–63.
Green, M. S. (2007) *Self-Expression*. Oxford: OUP.
Grice, P. (1969) 'Utterer's Meaning and Intention', *The Philosophical Review*, 78/2: 147–77. Repr. in Grice (1989), pp. 86–116.
——. (1989) *Studies in the Way of Words*. Harvard, MA: Harvard University Press.
——. (2001) *Aspects of Reason*. Oxford: OUP.
Harris, D. W. (2022) 'Semantics Without Semantic Content', *Mind and Language*, 37/3: 304–28. https://doi.org/10.1111/mila.12290.
Harris, D. W., Fogal, D. and Moss, M. (2018) 'Speech Acts: The Contemporary Theoretical Landscape', in D. Fogal, D. W. Harris and M. Moss (eds.) *New Work on Speech Acts*, 1–39. Oxford: OUP.
Heal, J. (2002) 'On First-Person Authority', *Proceedings of the Aristotelian Society*, 102/1: 1–19. https://doi.org/10.1111/j.0066-7372.2003.00040.x.
Hornsby, J. (2000) 'Feminism in Philosophy of Language: Communicative Speech Acts', in M. Fricker and J. Hornsby (eds.) *The Cambridge Companion to Feminism in Philosophy*, 87–106. Cambridge: CUP.
Horwich, P. (2005) *Reflections on Meaning*. Oxford: Clarendon Press.
Keiser, J. (2022) 'Language Without Information Exchange', *Mind and Language*, 37/1: 22–37. https://doi.org/10.1111/mila.12303..
Lepore, E. and Stone, M. (2015) *Imagination and Convention: Distinguishing Grammar and Inference in Language*. Oxford: OUP.
Lloyd, E. A. (2010) 'Confirmation and Robustness of Climate Models', *Philosophy of Science*, 77/5: 971–84. https://doi.org/10.1086/657427.
McDowell, J. H. (1980) 'Meaning, Communication, and Knowledge', in *Philosophical Subjects: Essays presented to P. F. Strawson*, 117–39. Clarendon Press. Repr. in McDowell (1998), pp. 29–50.

——. (1998) *Meaning, Knowledge, and Reality*. Harvard, MA: Harvard University Press.
Madva, A. (2016) 'Why Implicit Attitudes Are (Probably) Not Beliefs', *Synthese*, 193/8: 2659–84. https://doi.org/10.1007/s11229-015-0874-2.
Mandelbaum, E. (2016) 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias', *Noûs*, 50/3: 629–58.
Mele, A. R. (2001) *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.
Millikan, R. G. (2017) *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford: OUP.
Moran, R. (2005) 'Problems of Sincerity', *Proceedings of the Aristotelian Society*, 105/3: 341–61.
Neale, S. (1992) 'Paul Grice and the Philosophy of Language', *Linguistics and Philosophy*, 15/5: 509–59.
——. (2005) 'Pragmatism and Binding', in Z. G. Szabó (ed.) *Semantics versus Pragmatics*, 165–285. Oxford: Clarendon Press.
——. (2016) 'Silent reference', in G. Ostertag (ed.) *Meanings and Other Things: Themes from the Work of Stephen Schiffer*, 229–344. Oxford: OUP.
Pagin, P. (2011) 'Information and Assertoric Force', in J. Brown and H. Cappelen (eds.) *Assertion: New Philosophical Essays*. Oxford: OUP.
Quilty-Dunn, J. and Mandelbaum, E. (2018) 'Against Dispositionalism: Belief in Cognitive Science', *Philosophical Studies*, 175: 2353–72.
Reimer, M. (1992) 'Three Views of Demonstrative Reference', *Synthese*, 93/3: 373–402.
——. (2004) 'What Malapropisms Mean: A reply to Donald Davidson', *Erkenntnis*, 60/3: 317–34.
Ridge, M. (2006) 'Sincerity and Expressivism', *Philosophical Studies*, 131/2: 487–510.
Rosenthal, D. (1989) 'Postscript to "Intentionality"', in *Rerepresentations: Readings in the Philosophy of Mental Representation*, 341–344. Kluwer Academic Publishers. Repr. in Rosenthal (2005a), pp. 100–2.
——. (1998) 'Consciousness and its Expression', *Midwest Studies in Philosophy*, 22/1: 294–309. https://doi.org/10.1111/j.1475-4975.1998.tb00342.x.
——. (2005a) *Consciousness and Mind*. Oxford: OUP.
——. (2005b) 'Introspection and Self-Interpretation', in *Consciousness and Mind*, 282–305. Oxford: Clarendon Press.
——. (2005c) 'Why Are Verbally Expressed Thoughts Conscious?', in *Consciousness and Mind*, 282–305. Oxford: Clarendon Press.
Saul, J. M. (2012) *Lying, Misleading, and What Is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford: OUP.
Schiffer, S. (1972) *Meaning*. Oxford: OUP.
——. (1987) *Remnants of Meaning*. Cambridge, MA: MIT Press.
——. (1992) 'Belief Ascription', *The Journal of Philosophy*, 89/10: 499–521.
——. (2003) *The Things we Mean*. Oxford: Clarendon Press.
——. (2016) 'Gricean Semantics and Reference', in Ostertag, G. (ed.) *Meanings and Other Things:*, 493–527. Oxford: OUP.
Schupbach, J. N. (2016) 'Robustness Analysis as Explanatory Reasoning', *British Journal for the Philosophy of Science*, 69/1: 275–300. https://doi.org/10.1093/bjps/axw008.
Schwartz, J. (2020) 'Saying 'Thank You' and Meaning It', *Tandf: Australasian Journal of Philosophy*, 98/4: 718–31. https://doi.org/10.1080/00048402.2019.1698627.
Schwitzgebel, E. (2006) 'The Unreliability of Naive Introspection', *Philosophical Review*, 117/2: 245–73. https://doi.org/10.1215/00318108-2007-037.
Schwitzgebel, E. (2010) 'Acting Contrary to Our Professed Beliefs or the Gulf between Occurrent Judgment and Dispositional Belief', *Pacific Philosophical Quarterly*, 91/4: 531–53.
Scott-Phillips, T. (2015) *Speaking Our Minds: Why Human Communication is Different, and How Language Evolved to Make it Special*. London: Palgrave Macmillan.
Searle, J. (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: CUP.
Sellars, W. (1969) 'Language as Thought and as Communication', *Philosophy and Phenomenological Research*, 29/4: 506–27.
Simons, M. (2017a) 'Local Pragmatics in a Gricean Framework', *Inquiry: An Interdisciplinary Journal of Philosophy*, 60/5: 466–92. https://doi.org/10.1080/0020174X.2016.1246865.
——. (2017b) 'Local Pragmatics in a Gricean Framework, Revisited: Response to Three Commentaries', *Inquiry: An Interdisciplinary Journal of Philosophy*, 60/5: 539–68. https://doi.org/10.1080/0020174X.2017.1281205.

Sorokowski, P. *et al*. (2019) 'Voice of Authority: Professionals Lower their Vocal Frequencies when Giving Expert Advice', *Journal of Nonverbal Behavior*, 43/2: 257–69.
Sperber, D. and Wilson, D. (1986) *Relevance: Communication and Cognition*. Oxford: Blackwell.
——. (1995) *Relevance: Communication and Cognition 2 edn*. Oxford: Blackwell.
Stojnic, U. (2021) *Context and Coherence: The Logic and Grammar of Prominence*, Oxford, UK: OUP.
Stokke, A. (2014) 'Insincerity' , *Noûs*, 48/3: 496–520.
——. (2018) *Lying and Insincerity*. Oxford: OUP.
Unnsteinsson, E. (2017) 'A Gricean Theory of Malaprops', *Mind and Language*, 32/4: 446–62.
——. (2022a) 'The Social Epistemology of Introspection', *Mind and Language*, Early View. 1–18. https://doi.org/10.1111/mila.12438.
——. (2022b) *Talking About*. Oxford: OUP.
Weisberg, M. (2006) 'Robustness Analysis', *Philosophy of Science*, 73/5: 730–42. https://doi.org/10.1086/518628.
Wharton, T. (2009) *Pragmatics and Non-Verbal Communication*. Oxford: CUP.
Williams, B. (2002) *Truth and Truthfulness: An Essay in Genealogy*. Princeton, NJ: Princeton University Press.
Wilson, D. and Sperber, D. (2012) *Meaning and Relevance*. Oxford: CUP.
Wilson, T. D. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.
Woodward, J. (2006) 'Some Varieties of Robustness', *Journal of Economic Methodology*, 13/2: 219–40. https://doi.org/10.1080/13501780600733376.

[1] *University College Dublin, Ireland*
[2] *University of Iceland, Iceland*