Received: 9 March 2021 | Revised: 13 January 2022 | Accepted: 17 February 2022

DOI: 10.1111/mila.12438

SUBMITTED ARTICLE

WILEY

The social epistemology of introspection

Elmar Unnsteinsson[1,2]

[1]School of Philosophy, University College Dublin, Dublin, Ireland

[2]School of Humanities, University of Iceland, Reykjavík, Iceland

Correspondence
Elmar Unnsteinsson, School of Philosophy, University College Dublin, Newman Building, Office 502, Dublin, Ireland.
Email: elmar.unnsteinsson@ucd.ie

Funding information
Icelandic Centre for Research, Grant/Award Number: 206551

I argue that introspection recruits the same mental mechanism as that which is required for the production of ordinary speech acts. In introspection, in effect, we intentionally tell ourselves that we are in some mental state, aiming thereby to produce belief about that state in ourselves. On one popular view of speech acts, however, this is precisely what speakers do when speaking to others. On this basis, I argue that every bias discovered by social epistemology applies to introspection and other forms of self-directed representation. If so, it becomes unclear in what sense social epistemology is social.

KEYWORDS
Gricean pragmatics, insincerity, introspection, self-deception, social epistemology

1 | INTRODUCTION

Why do we sometimes tell ourselves that we believe something or other? More precisely, what are our reasons for doing so and what is the mental mechanism in virtue of which we can engage in such acts of inner representation? In this article, I argue that there is no essential difference between self-directed speech and other-directed speech at the levels of mental mechanism, normal function or, even, psychological motivation. Roughly, we tell others that *p* to try to activate some *p*-attitude in their minds. This is an intentional action whose competent performance is most likely explained by the operation of various mechanisms geared toward the production of evidence intended to be interpreted by others who share the same competence. And, even when engaged in the process of introspection—to ascertain our own attitudes and mental

states—we perform intentional acts of exactly this kind. Most importantly, our goal is to produce beneficial cognitive effects of some sort in ourselves.

As an empirical hypothesis, this may sound perfectly plausible. In this article, however, I try to show that it is more than a mere hypothesis by pulling together strings from speech act theory, the cognitive science of introspection and inner speech, as well as the literature on self-deception. Taking a glimpse into each of these areas makes it possible to paint a picture of mental states and the ways in which we represent those states to ourselves and others in action. This picture suggests that self-directed representation and other-directed representation are far more alike than common sense, and a number of theorists, seem to suppose.

Finally, I point out that this conclusion will have resounding consequences for social epistemology and related fields, because every single phenomenon of interest—bias, lying, testimonial injustice, echo chambers, and so on—to theorists working in that field should have direct analogs in the study of self-directed speech. This opens up a can of worms. But, it is a can of worms well worth serious attention, because any malignant purpose we are afraid others might have in talking to us, we are just as likely to have ourselves in the privacy of our own minds. In some sense, of course, this is a pessimistic view of our introspective capacities, but that is not really my point here. The point of interest is that social epistemology has been too focused on various doxastic or epistemic effects of informational interactions between agents, to the detriment of studying the very same effects as they manifest themselves in reasoning and cognition themselves. Furthermore, if the account offered here is right, introspection is a thoroughly social phenomenon, because it recruits cognitive mechanisms designed to facilitate social interaction. So, the article is not exclusively about social epistemology, but also the very sociality of introspection.

Here is the plan. In Section 2, I explain and motivate a distinction between representational states (like propositional attitudes) and representational acts. The acts have representational import partly in virtue of being grounded in some representational state and, further, are intentionally performed to have cognitive effects on other minded creatures. If so, I argue, we should not assume that the mental mechanisms of self-directed representation is any different from other-directed mechanisms. In Section 3, I argue that introspection essentially involves self-directed representational acts of this same sort, most likely recruiting the same mechanism. In Section 4, I show that our motivations for deceiving others are not essentially different from our motivations for deceiving ourselves. In particular, we are often motivated to deceive ourselves to be better at deceiving others. Finally, in Section 5, I put the pieces together to paint a picture according to which the results of social epistemology, and social psychology, can be applied directly to self-directed acts of representation. The moral of the story is, roughly, that social epistemology and related fields suffer from *interpersonal* or *interactional* bias.

2 | ATTITUDES AND REPRESENTATIONAL ACTS

There is a fundamental distinction between *someone believing p* and *someone performing an action whereby they represent themselves as a p-believer*. The first is some state of the believer and the second an action performed by the believer. Let us call the first an *attitude* and the latter *representational acts* or, to speak more generally, representational dispositions. So, when I believe *p* but insincerely tell you *not p*, there is a mismatch between my attitude and my representational act; I have a disposition not to represent myself as a *p*-believer *to you*, while in fact I am a *p*-believer. I say that this distinction is fundamental because, while the acts are under

considerable intentional control, attitudes are widely thought to be much less so. When I see an apple in front of me, the normal functioning of the mechanisms of perception and belief-fixation will typically make it so that I believe that there is an apple in front of me. Staring at the apple, I am normally quite unable to make myself believe that it is not an apple. Nothing could be further from the truth when it comes to the corresponding representational act or disposition. It is part and parcel of my cognitive competence that I can easily say, to myself or others, that I do not believe there is an apple in front of me.[1]

Let us consider representational dispositions in more detail, however. Dispositions to represent oneself as having some contentful mental state, like the belief that *p*, are rational dispositions of intellectually sophisticated creatures with the capacity to guide their actions toward the achievement of some perceived goal. We can readily accept that nonhuman animals might have various mental attitudes, but it is less clear whether many of them have such dispositions. So, let us focus on humans. If I decide, on a given occasion, to perform some action whereby I represent myself as a *p*-believer, I will normally have some reason to do so. The reason typically consists in some intention of mine, directed toward the achievement of some effect which I perceive as beneficial. Of course, the act may be performed out of mere habit or without any conscious reflection. But the type of action at issue is, still, a paradigm case of actions needing considerable cognitive sophistication, even if effort and preparation can be significantly reduced in various ways on particular occasions.

In general, what would be a reasonable expected effect, perceived as beneficial, of performing these kinds of representational acts? Well, first of all, the effect must be envisaged as taking place in a similarly sophisticated cognitive system, that is, a system which can be at the receiving end of acts of representation. Disregarding Artificial Intelligence-systems and possible communication with animals or aliens, the expected effect will have to be a cognitive effect in some human being. At the very least, I will know that the act will not have any effect, beneficial to myself, on bananas or muddy puddles. The point here is, simply, that representational acts are a paradigm case of actions studied by decision theory, ultimately explained in terms of practical reason and action-controlling intentional states, directed at someone who is also cognitively equipped to make decisions and inferences.

It is a reasonable hypothesis, then, that these acts correspond roughly to what Gricean intentionalists would call acts of speaker meaning.[2] Consider a typical Gricean theory of acts of telling someone that *p*. In such a case, the speaker S must perform the act of uttering something X, with the intention of thereby producing or activating in some addressee H the belief that *p*. What S desires to achieve, then, is a specific cognitive effect in H, and it is easy to see how this might be perceived to be beneficial in many cases. S forms the belief that uttering X, which is perhaps some sentence in English, in H's presence, is the most efficient way, in context, to achieve the desired effect.

Let us call speakers' competence to perform and interpret acts of speaker meaning "pragmatic competence". A plausible theory of pragmatic competence will specify the cognitive

[1]Note that the distinction is between first-order attitudes or states and *acts* representing those attitudes or states. Thus the distinction is similar to, but still importantly different from, the one employed by higher-order theorists of consciousness (e.g., Rosenthal, 2002). The representational act seems to be second-order, in some sense, to the state represented. But it is not a representational state representing another representational state, it is merely an act. I go into more detail about the act-state distinction in Unnsteinsson (2022, Chapter 3).

[2]See, for example, Grice (1989, 2001), Bach and Harnish (1979), Carston (2002), Harris (2021), Neale (1992, 2005, 2016); Schiffer (1972, 1987, 2003), Scott-Phillips (2015), Simons (2017a, 2017b), Sperber and Wilson (1986/1995), and Wilson and Sperber (2012).

mechanisms in virtue of which speakers can perform these acts. Cognitive mechanisms, like memory or perception, are always assumed, implicitly or explicitly, to have some function or other. Functional hypotheses of this sort are necessary to guide the construction of explanations and theories, if nothing else, and are themselves subject to revision like any other hypothesis. The function of a cognitive mechanism is specified by its characteristic effect.[3] The Gricean intentionalist, as I understand the idea, hypothesizes that the characteristic effects of the mechanism in virtue of which speakers have pragmatic competence is, simply, the range of effects speakers intend to have on addressees in performing acts of representation or communication. This, finally, affords a notion of normal function. The mechanism in question performs its normal or proper function exactly when its operation successfully produces the characteristic effect (for more detail, see Unnsteinsson, 2022).

This is of course highly schematic. So, let us fill in some of the blanks. The mechanism of pragmatic competence, on this story, is partly constituted by the operation and interaction of sub-mechanisms, which we can specify as information-carrying states with specific functions, but not in terms of neurobiology, or at least not yet. The first is what I will call the sub-mechanism responsible for the capacity to form *effective intentions*. This is simply an intention to produce a cognitive effect in someone, for example to produce a belief or intention with some specific content. The second is what I will call the *signaling intention*. This is the speaker's intention to produce in someone the *recognition* of the effective intention. On some views, there will also be the intention that the hearer's satisfaction of the signaling intention constitute a reason to actually satisfy the effective intention. This is sometimes called the *directive* or *Gricean* intention, but it can be safely ignored here. The combination of these three, or only the first two, we can call the *communicative intention*. There is a lot of detail, which I leave out here, but this is the basic story about the mechanisms postulated by Gricean intentionalists to explain the characteristic effects of pragmatic competence. The speaker utters something, for example to *tell H that p*, only if they have a communicative intention embedding an effective intention whose *effect* is supposed to be *H believing p*.[4]

When postulating mental mechanisms to explain some capacity, it is sound methodology to focus, in the first instance, on capacities universally attested in the relevant species. Speech act theory has traditionally focused on so-called speech act verbs in English and, sometimes, taken highly institutionalized and potentially culture-bound actions like promising or apologizing as their paradigms (Austin, 1975; Searle, 1969). We should expect, rather, that whatever we postulate must ultimately explain the capacity to perform speech acts of promising, but also the whole gamut of promise-like acts human beings have performed from the beginning, and will perform in the future, in different languages and societies. The Gricean story is an hypothesis of this sort, because differences in speech act can be determined by differences in the effect embedded by the effective intention. For example, assertives will produce belief-like states, directives intention-like states, and questions inquisitive states. Combinations and variations of these basic speech act types will go an awfully long way to explain more sophisticated language-involving act-types.

[3]On mechanistic explanation in cognitive science, see Bechtel (2007), Craver (2007), and Glennan (2017). On the notion of function in mechanistic explanation, see Garson (2013, 2017, 2022) and Neander (2017). For more discussion on the relevance of both to pragmatic competence, see Unnsteinsson (2022).

[4]This is a slight modification of the standard terminology in relevance theory (Carston, 2002; Sperber & Wilson, 1986; Sperber & Wilson, 1995; Wilson & Sperber, 2012), where "effective" is "informative" and "signaling" is "communicative". I want to preserve "communicative" for the whole and I think "effective" is better because many utterances are not primarily intended to impart information.

I take these considerations together as justification for taking the Gricean functional hypothesis seriously. That is to say, there is a compelling case for thinking that humans possess so-called pragmatic competence in virtue of the benefits of being able, quickly and efficiently, to have cognitive effects on other humans. The added structure of signaling intentions—namely, letting others know that you have some effective intention or other—also has obvious benefits to individuals and the species as a whole. The development of a shared, spoken language can then be thought of as, so to speak, the "Cambrian" explosion of expressibility, because it adds enormously to the efficiency of communicative episodes, reducing cognitive load at the same time. There are other functional hypotheses, of course, the most obvious being that language-involving capacities, recursion in particular, function to enable humanlike thought (see, e.g., Chomsky, 1975; Hauser et al., 2002). As I see it, there is no incompatibility here; both can be right in different ways. But that is controversial territory I will have to shy away from for the moment. So, the upshot is merely that pragmatic competence as here described is well motivated as a universal or near-universal human capacity and I make no particular commitment about whether this function was selected-for by a process of evolution by natural selection or, for example, whether it is really a by-product of the evolution of the language faculty (which is there for different reasons). Crucially, however, to assign the capacity to humans is to postulate cognitive mechanisms in virtue of which they can form *intentions* to have *cognitive effects* on *minded creatures*.

This account now needs to be more firmly connected to the distinction between attitude and representational disposition. First, we should generalize the latter so as to include all and only those acts whereby agents represent themselves as having some *mental attitude or state*, not merely belief. Second, we should restrict our theoretical attention to the class of actions under intentional control, for reasons already discussed. Thirdly, it is reasonable to accept the Gricean story as a working explanatory hypothesis whose explanandum is precisely the human capacity to perform representational acts, so understood, intentionally. The story is incomplete, but it is one that would, if true, help to explain the human capacity to perform representational acts. Finally, this explanation assumes something like the distinction between attitudes and acts; we need a prior account of the attitudes to serve as the effects embedded by the effective intention in action. Moreover, the account assumes that there are real mental attitudes classified as intentions, effective intentions, signaling intentions, and so on. The method here is to assume that states confer contents on actions, if the latter are to be contentful at all, in virtue of some causal relationship.

We have then, with good reason I believe, eliminated from our explanatory purview any intentional, representational action whose aim is not one of somehow influencing the mental or cognitive state of a possible interpreter. This is not to say such actions are impossible or not interesting, only that they would constitute a malfunction or abnormal function of the postulated mechanism. For example, if the mechanism in question is genuinely put into operation, on a given occasion, to influence the mental contents of a muddy puddle, especially if the agent knows full well that the puddle will not be affected in any relevant manner, the mechanism is not serving its normal function on that occasion.

Now we can finally consider a basic distinction between possible interpreters. This is the distinction between self and other. Some representational acts are other-directed in that they are intended to influence the mental attitudes of someone other than the agent or speaker. But, of course, people often engage in various forms of *inner* or *private* speech, the nature and function of which can appear more mysterious. There is a long tradition, in the philosophy of language, of theorists who think that even if something like the Gricean mechanism is plausible in

the case of other-directed representational acts, this simply cannot be so for self-directed inner acts.

Surely, self-directed speech can be puzzling but much of it is quite mundane and amenable to simple, everyday explanation. Why did I say "Cheese!" aloud to myself in the supermarket yesterday? Well, I was reminding myself not to forget to buy cheese and, so, my intention was to have a cognitive effect on an audience, namely myself. Certainly, vocalization was strictly unnecessary. I could have reminded myself by saying the same thing silently. Already this presents a small puzzle. If I am to form the intention to remind S that *p* I must believe that S needs to be reminded that *p*. But if I am identical to S it seems that S does not need to be reminded that *p*, since to form the intention to remind someone that *p* requires that one already remember that *p* oneself. And so I cannot, in that case, really believe that S needs to be reminded that *p*. The form of the puzzle ought to be familiar from debates about the nature of self-deception—Alfred Mele (2001, p. 8) calls it the dynamic puzzle—but, as this shows, it can be generated merely by considering self-directed acts of representation.

If the dynamic puzzle can thus apply much more broadly than to acts of self-deception, we have less reason to suppose that the puzzle itself is a reason to believe that the phenomenon described does not exist. That is to say, self-directed inner speech, even if explained by the Gricean mechanism, is more credibly postulated than self-directed acts of intentional deception. There are two reasons for this. First, the phenomenon is simply all too familiar; we definitely voice our own attitudes to ourselves, knowing full well that we have those attitudes ourselves, and yet our motivation appears to be one of somehow influencing our own attitudes. A simple example, which would explain why I reminded myself about the cheese while avoiding the dynamic puzzle, is when I am motivated to broadcast the attitude more widely in my own mind. Sometimes the mere verbal articulation, self-directed, of some belief or intention, serves more robustly to activate or stabilize that very belief or intention and, thereby, to help with the intentional control of overt bodily action for example. Thus, on this way of thinking, we are saved from puzzles by ever so slight differences in the strength or nature of the *attitude* rather than in the proposition within its scope, before and after the self-directed representational act.

At the same time, it is important to note that self-addressed speech acts do not automatically produce the attitudes intended. For one thing, this would amount to doxastic voluntarism, which is not widely accepted. If I wanted to believe *p*, all I would need to do would be tell myself that *p*. For this reason, the distinction between the hearer's satisfaction of the signaling intention and the hearer's satisfaction of the effective intention is relevant in explaining both self-directed and other-directed speech acts. Take self-addressed imperatives, for example. I may recognize my own effective intention in telling myself to stop smoking, even if I do not thereby form the intention to stop smoking. My utterance may be to the benefit of other participants in the conversation, even if it is self-addressed, partly aimed at making them believe, falsely, that I do intend to stop. In such cases, I satisfy my own signaling intention without satisfying the effective intention. But the details of this account will have to await another occasion.

It is worth mentioning, in this context, that in other-directed assertoric speech, some theorists would argue that there is a fundamental divergence of interest between the speaker and the addressee (Sperber, 2013; Sperber et al., 2010). That is to say, the addressee benefits from the speaker's truthfulness, while the speaker benefits from the addressee's trust. But how could this divergence of interest make sense when the speaker is identical to the addressee?[5] Well, even if my interest *as a parent* may diverge from my interest *as a teacher*, some acts may be

[5]Thanks to an anonymous reviewer for this journal for raising this question.

intended to benefit me in both roles. Similarly, a self-addressed speech act will tend to benefit me both as speaker and addressee if I am truthful and trusting. By analogy, there is a sense in which it tends to benefit two communicators *as a group* that the speaker is truthful and the hearer trusting. In self-talk, if I am truthful but not self-trusting I will not believe the truths I tell myself. If I am not truthful but self-trusting I will believe the falsehoods I tell myself. Part of my point in this article, however, is that we sometimes go against our best interest, even in self-talk. Thus, we may deceive ourselves.

The second reason why we should explain inner speech within a Gricean framework is more significant, however. Recent work in empirical psychology and cognitive science shows that there is considerable variety in the nature and function of the class of actions normally labeled as inner speech. Let us consider three important distinctions. One, there is the distinction between self-directed speech and merely imagined speech, well articulated by Daniel Gregory (2016). Sometimes we imagine speech, for example when rehearsing a public talk, but at other times we explicitly address ourselves, for example in telling oneself that the talk should not start like that or some such. Two, some speech is inner in the sense that it is sub-vocal, but at other times it is merely private in the sense that there is no one around to hear what we say out loud to ourselves. Third, some inner speech is self-attributed and some is not self-attributed, which gives rise to a popular explanation of some cases of *auditory verbal hallucination* (AVH). On this way of thinking AVH occurs when episodes of either self-directed or imagined speech, produced and interpreted by a single speaker S, fail to be self-attributed by S. All of these categories are cross-cutting and there are many others worth our theoretical attention. I recommend Langland-Hassan and Vicente's (2018) collected volume on *Inner speech* for a comprehensive perspective on the topic.

The only point to be made here, and this is finally the second reason, is that within the heterogeneity of types we can carve out a natural place for inner speech which is literally self-directed. This category itself is needed to explain some cases of AVH, but also to explain the occurrence of self-directed reminders, pep talk, or inner criticism. But seeing that the category of literally self-directed inner speech, where "literally" is supposed to evoke the idea that it is explained by the normal function of the Gricean mechanism, falls into place within a much broader spectrum of capacities, actions, and occurrences, helps to dispel the initial puzzlement. Because much of inner speech can be seen to have easily understandable functions, which may or may not be explanatorily posterior to the Gricean mechanism, like that of planning, monitoring, and rehearsing behavior, controlling emotions, and cultivating creativity (e.g., Alderson-Day & Fernyhough, 2015). The same basic point applies to some types of other-directed representational acts, for example, the class of non-intentional behaviors which still carry some information, like a cry directly caused by an injury. Strictly speaking, these will be abnormal functions of the Gricean mechanism or, more likely, simply the operation of a different but related mechanism. The argument in this article does not require any specific commitment on which mechanism is prior in ontogeny, phylogeny, or otherwise. The only conclusion to be drawn is that the Gricean mechanism is plausibly postulated to explain an important range of representational acts, both when self-directed and when other-directed. That is to say, there need not be any essential difference between the normal function, mental mechanism, or cognitive competence in virtue of which speakers perform self-directed and other-directed representational acts, as here understood.

Finally, it should be noted that the literature on inner speech is replete with theories or data to the effect that are substantial and interesting analogies between self-directed and other-directed speech, both in production and interpretation (e.g., Carruthers, 2011, 2018;

Fernyhough, 2016; Levelt, 1989; Postma, 2000; Vicente & Jorba, 2017). As far as I can tell, however, none goes so far as to apply, without any modification, the Gricean mechanism of interpersonal pragmatic competence to the intrapersonal case. Even more, theorists who emphasize the analogies in their accounts have sometimes explicitly rejected any such extension. Peter Carruthers, whose interpretive theory of inner speech and introspection are major inspirations for many of the ideas in this article, argued recently that "in inner speech there generally *is* no communicative intent" (2018, p. 46, emphasis his). I plan to disagree.

3 | WHAT IS INTROSPECTION?

In this section, I argue that any plausible theory of introspection will presuppose that it partly consists in the intentional performance of self-directed representational acts. More precisely, any episode of introspection will involve agents in some kind of act whereby they indicate to themselves that they have some mental state or other. I switch to "indicate" here because of the theoretical baggage often associated with the term "representation". Many theorists would refrain from committing to the existence of mental representations literally stored in the mind but, luckily, that controversy can be sidestepped. It is assumed here that speakers have attitude-states, but these can be theorized as dispositional or even interpretationist phenomena and, moreover, the corresponding representational act merely purports to carry or convey the information that some specific such state is realized. It is possible that the information is carried by the act without there being, literally, a mental representation the content of which explains the informational content of the act. I will continue to talk in terms of "representational" acts, however, with this caveat in place.

By way of generalization, philosophical theories of introspective processes (IPs) or episodes seem to suggest a mechanism, which decomposes into four parts. For a helpful overview of these theories, see Carruthers (2011) and Schwitzgebel (2019), to which I am indebted, although the following account departs from theirs in some ways. Typically, IPs are theorized as proceeding through four steps; TARGET, SENSITIVITY, ENACTMENT, and EFFECT:

1. TARGET: First-person mental state M.
2. SENSITIVITY: Attention to or detection of M.
3. ENACTMENT: Judgment that M obtains.
4. EFFECT: Belief that M obtains.

First, IPs will target some first-person mental state or phenomenon, like an attitude or an emotion. The target is, simply, that about which the IP is supposed to produce belief or knowledge. For instance, let us say I have a mental state of liking Bach at some time *t*. The state is active in the minimal sense that it is ready to exert causal influence on my behavior if the opportunity arises. Any IP worthy of the name will target some such state or other.

Even this may sound a bit contentious. Some theories of IPs can be thought of non-cognitivist and others as cognitivist and the former can be described, sometimes at least, as denying the idea that there must be some prior state the existence of which introspection aims to discover. I have in mind expressivist views like Bar-On (2004) or self-fulfilling views like Moran (2001). On these theories, IPs may have a role to play in the very production of these states, not in their detection. A more fruitful way to look at this is as follows. We should assume that different theories can differ on the initiation and termination conditions of IPs, and some

may even insist that one or more of the four steps is optional. So-called noncognitivist theories, then, may hold that IPs are really initiated at step 3 and, perhaps, terminated at step 1 (or 2). In this way, my judgment that I like Bach might produce my liking of Bach. Target becomes effect.

The next step, SENSITIVITY, seems potentially optional. Many so-called noncognitivists would argue that this step does not exist, for assuming otherwise seems to suggest that there is a faculty of inner perception charged with detecting mental states. Moreover, it implies that TARGETS are real mental phenomena about which IPs can produce knowledge. Both would be overreactions. The detection in question could be performed by the very same cognitive mechanism responsible for the detection of mental states which are not first-personal (see, e.g., Carruthers, 2011). Further, most noncognitivists about introspection will be realists about the TARGET even if they happen to think they are produced rather than detected by IPs. Anyway, as we have already seen, anti-realism (e.g., interpretationism) about first-person mental states themselves is beside the point. Still, we should allow logical space for theories on which IPs run from ENACTMENTS to TARGETS, skipping SENSITIVITY altogether.

For my argument, ENACTMENT is the single most crucial step in the introspective process. The basic point here is that introspection must name some type of mental activity whereby the actor aims to produce or detect some specific mental state rather than another. Of course, the specificity itself can vary and is a matter of degree. I will assume that the mental act of judgment is a plausible description for this type of activity and the act's EFFECT is belief or knowledge about the state. At any point in time, presumably, the thinker is in more than one mental state (or TARGET) and, in principle, both SENSITIVITY and ENACTMENT can selectively attend to one rather than the other. Perhaps SENSITIVITY tracks all TARGETS at the same time. But if there is anything special about IPs, it will be the selectivity of the judging act. If the thinker is currently in two states, M1 and M2, they are free to select only M1 as the state judged to obtain. The multiplicity of available first-person mental states mandates some measure of choice and judgment, that is, the intentional performance of a mental act, in the introspective process.

Note that this makes introspection different from consciousness, even if the latter is thought of as an automatic feature of some mental states. Take, for example, Uriah Kriegel's self-representationalism (Kriegel, 2009). On his view, a conscious state is conscious in virtue of representing *itself*, besides anything else it might represent. Even if this is right, it need not follow that any conscious state is introspected, because introspection involves intentionally controlled selection—possibly between two *conscious* states, M1 and M2—in the service of producing some further doxastic or epistemic upshot.

The argument here is emphatically not that the belief that M1 obtains cannot come into existence in any other way, that is without some intermediate act of judgment. It is entirely possible that being in a first-person mental state M1, or mere attention to that fact, could produce the belief that one is in state M1. This would just not constitute a plausible theory of *IP*s, for it makes introspection automatic, effortless, and (more troublingly) exhaustively iterative. Certainly, belief-formation is very often automatic and effortless, but it is not exhaustively iterative at the same time. To see this, imagine a theory of IPs on which nothing like ENACTMENT is needed to connect the TARGET and the EFFECT (similar, perhaps, to Lycan, 1996). Since every EFFECT is also a TARGET, because the belief that M obtains is a first-person mental state, the process would spiral endlessly, clogging the mind with vast collections of redundant doxastic states. Surely, then, some filtering mechanism is necessary, to select which states are targeted by IPs, and this is the crucial role performed by ENACTMENT.

Now, if the act of judging that one is in mental state M1, when one is also in M2, is an intentional mental act it is plausible to think, as we argued above, that it is performed with an eye to some beneficial cognitive effect predicted by the agent. Most theories suppose that the upshot of a successful IP is some doxastic or epistemic state, whereby one comes to believe or know something about one's own mental state. We will go along with this assumption here, noting only that this certainly does not preclude such acts from having additional effects, intended or otherwise. Such additional effects will become important later. So-called noncognitivist theories would, again, insist that IPs are or can be self-fulfilling or self-constituting, such that the process terminates in the production of the TARGET, rather than beliefs about the TARGET.

Finally, using the example of *liking Bach*, let us spell out one possible IP.

BACH

1. TARGET: I like Bach.
2. SENSITIVITY: I attend to or detect my liking of Bach.
3. ENACTMENT: I judge that I like Bach.
4. EFFECT: I believe that I like Bach.

As an introspective process, we should think of my initial state as one where I like Bach without actively believing that I like Bach. Otherwise, BACH is superfluous. It might be an unreflective, simple fact about me that I always enjoy myself when I hear Bach's music. Reflectively, I may be just as likely to think of myself as liking Schubert, which might also be true. There is some mechanism of detection by which I attend to my own state of liking Bach. Finally, I perform the act of judging that I like Bach and, if the terminal state of an IP is supposed to be the formation of a relevant doxastic state—like the one in EFFECT—part of my reason or motivation for so judging on that occasion is to produce or activate that state. Notice, also, that the operation of the mechanism may result in the formation of a number of other doxastic states, for example the belief that I intend to produce a specific belief in myself, but these are not thereby introspective beliefs. Introspective beliefs must be the TARGETS, and the objects of SENSITIVITY, in a particular introspective process.

Crucially, however, every cognitive mechanism can be dysfunctional on a given occasion of operation. So, for example, I might undergo ENACTMENT to produce the EFFECT without success. Immediately after, I might judge that I like Schubert more and fail to activate the belief that I like Bach. More obviously, perhaps, it might simply be false that I like Bach; I performed the act of judgment merely because the people around me tend to like Bach and I really want to be more like those people. In fact, Bach makes me nauseous, even if when listening I am not aware of the composer's identity.

This account of introspection is merely a generalization of many different philosophical theories of the phenomenon. It is not uncontroversial, but still it is flexible enough to accommodate differences in emphasis between theories. Now I will argue that if the generalization is accepted we can conclude that the third step in introspection, that is, ENACTMENT, can be identified as a self-directed representational act, the aim of which is to produce or activate some specific cognitive effect in oneself. In BACH, and depending on the theory, the effect is going to be the belief that I like Bach, or the very fact that I like Bach, or perhaps both.

To illustrate the point, imagine Peg undergoing BACH in the company of her old friend Greg. Reflecting quietly on her taste in music and her reactions to different composers, she gradually comes to realize that she enjoys Bach. The realization prompts her to judge that this

is so and, knowing that Greg would greatly appreciate her new-found sophistication, she judges by performing a speech act, uttering:

(1) I like Bach.

Peg utters (1) with an intonation and emphasis which is intended to signal certain specific characteristics to Greg, namely that her enthusiasm for Bach came to her as a personal revelation, based on inner reflection and deliberation about her tastes and distastes.

Now note that Greg's presence in this illustration is not necessary to explain the behavioral implementation of Peg's act of judgment. It would have been rational for her to utter (1) only to herself, either out loud or sub-vocally. This is merely one of the ways in which she could choose to implement the act in question. Neither is it necessary that judgments or representational acts more generally are realized in language, manifested internally or externally. The judgment could be realized in an iconic medium, like a mental image, rather than in any language-like discursive medium. But the representational implementation of an intentional action need not make any difference to the goal or expected effect of performing the action. When I want to cheer myself up, I may either decide to meet with friends or to think happy thoughts. Both actions can be different ways to achieve the same intended goal: cheering myself up. Similarly, saying (1) in inner speech could be one way for Peg to judge that she likes Bach, but there are others. ENACTMENT is not defined in terms of any particular representational medium; it is an act intended to produce doxastic EFFECTS in oneself.[6]

Plausibly, then, the effect Peg intends to produce by uttering (1) is the same in both situations, only directed at different cognitive agents. With Greg around, Peg is trying to produce the belief that she, Peg, believes that she likes Bach, in Greg. With Greg gone, she is trying to produce that very same belief in herself. Also, presumably, Greg's presence does not mean that Peg cannot intend to produce the belief in herself as well, engaging in a partially open process of introspection, for Greg to see.

It follows that ENACTMENT-steps, as they occur in IPs, constitute representational acts. In judging that I like Bach, I perform an action whereby I represent myself as liking Bach. If the IP is performed privately, not to the benefit of anyone other than myself, my act is intended to represent myself *to myself* as liking Bach. I intend to categorize myself as one of the Bach-likers, for my own benefit. Thus, if representational acts are to be identified with Gricean acts of speaker meaning, the ENACTMENT-step is constituted by an act performed with a communicative intention, only self-directed. This is not surprising, for we have already identified all of the necessary elements of such intentions in IPs. First, there is a purported first-order mental state (*liking Bach*, the TARGET). Second, there is an intentional action the performance of which is supposed to represent or indicate the presence of this first-order mental state (ENACTMENT). Third, there is some cognitive EFFECT which is intended as the primary or most immediate consequence of the act's performance. We have hypothesized, along with many theories of IPs, that the consequence is some relevant doxastic or epistemic state. It seems eminently plausible to

[6]This point about implementation may appear to suggest that, possibly, the argument of this paper could be constructed without any reference to speech acts, whether inner or outer. That is, we could make do with any behavioral implementation whatsoever, if it has the right function relative to a given IP. Strictly speaking, that is true. But the point here is that the structure of IPs is *communicative*, and equivalent to the structure of the interpersonal case. And speech acts, according to the Gricean framework, are best thought of as *communicative acts*; any behavior intended as evidence for a communicative intention (cf., Grice, 1989, p. 92, on his "artificially extended" use of the word "utterance"). Thanks to a reviewer for this journal for discussion on this point.

suppose that introspection is partly constituted by the performance of an act of self-communication. If this is correct, it becomes very easy to understand why IPs can fail to deliver the truth. We are obviously motivated, often unconsciously, to misrepresent our attitudes to others. But we are also, and often for the same reasons, motivated to misrepresent our attitudes to ourselves. This is the topic of the next section.

4 | DECEIVING ONESELF THE BETTER TO DECEIVE OTHERS

Why would we misrepresent our own mental states or attitudes to ourselves? If we accept only part of the evidence hailing from work in social and cognitive psychology, we can find plenty of reasons. First, we might simply not know, and have no easy way of finding out, the truth about our own mental state and, thus, any self-directed representational act will risk misrepresentation. Second, and more importantly for my purposes, we can lie to ourselves for all of the same reasons we have for lying to others. Most people desire to sustain some particular self-image—for example, as competent, honest, deserving, and so on—not only in the eyes of others, but also in their own. In this section, I argue that self-deception is a ubiquitous phenomenon and, second, that sometimes self-deception is actually motivated by the thinker's desire to get away with deceiving others. This is evidence that, even in basic motivation, self-directed and other-directed acts of misrepresentation have a common causal structure.

Research in social psychology is taken by many to support the first claim as well as supporting the idea, needed for the second, that self-deception can be nonconscious yet strategic, goal-directed, and flexible (Carruthers, 2011; Funkhouser & Barrett, 2016; Mercier & Sperber, 2017; Trivers, 2011; Wilson, 2002). Unfortunately, however, as Eric Funkhouser (2017) points out, it is common in psychology to understand self-deception very broadly, to include almost any type of bias in the formation or perpetuation of contentful attitudes. Funkhouser (2017, pp. 223–224) thus proposes to distinguish *self-delusion* and *self-deception*. Self-delusion produces false beliefs outright without the thinker retaining any awareness at all of the actual truth. And the false belief is intuitively explained in terms of a thinker's motivational states, for example, the desire to believe one is more attractive than one really is. When self-delusion is perfect and complete there is, then, no internal conflict in the subject's mind. But conflict is essential to self-deception. The self-deceived person retains some inkling of the truth and this unconscious awareness sometimes guides their action or inaction. Funkhouser uses the example of a bald man who does not want to believe he is bald although, of course, deep down he cannot help knowing. When the topic comes up in conversation he tries to avoid it and if asked he will be disposed to deny being bald. He also tends to avoid looking into mirrors. The self-deception is also explained in terms of the thinker's motivational states, in this case his desire not to be bald or, perhaps, not to believe he is bald (Funkhouser, 2005).

Much of the evidence hailing from social psychologists can, it seems, only be taken to support the prevalence of self-delusion. Confirmation bias, for example, need not produce internally conflicted, self-deceived individuals, since it is simply a general strategy of seeking information in such a way that the beliefs one already has are more likely to be retained than replaced by some new beliefs. Properly understood, however, the research can also be taken to support the idea that self-deception is common in humans. This is simply because many cases of self-delusion originate in or naturally give rise to corresponding self-deceptive states. The bald man might possibly be deluded, believing flat-out that he is not bald, but this would

require radical departures from normal processes of belief-fixation, as the evidence to the contrary is ever present and almost undeniable. So, this is a case where self-deception is much more likely to occur than self-delusion.

Take so-called attractiveness bias as a second example. According to some empirical work, people are generally prone to believe that they are 20% more physically attractive than they "actually" are (Trivers, 2011, p. 16). This is taken to be shown by people's stronger tendency to identify themselves in pictures of themselves that have been tampered with to enhance known features of perceived attractiveness (e.g., symmetry) than with pictures that do the opposite. Clearly, such a bias need not result in internal conflict; we will just believe flat-out that we are prettier than we in fact are. Having this belief could be motivated by increased confidence and a better self-image (Funkhouser, 2017, p. 233). Still the bias also helps to explain why people sometimes experience episodes of self-deception about their own attractiveness. Looking at yourself in the mirror you notice a newly formed indication of old age—wrinkle, grey hair, bald spot, whatever—this is evidence you cannot ignore that, by some societal standard you have internalized, you are less attractive than before. Slowly but surely, you start convincing yourself that you are still attractive, even that these features add to your attractiveness in some way. The result is a classic episode of self-deception and, if that is right, the internal conflict essential to self-deception is indeed deeply rooted in human psychology and should be quite common. All we need to add to known cases of biased self-delusion is the accessibility of clear evidence contradicting the delusional belief and we predict, at least, conflict-ridden episodes of self-deception. The result of those episodes may well be self-delusion rather than semi-permanent mental fragmentation, but the point still holds.

Funkhouser's distinction is important to the dialectic of this article. As pointed out in Section 2, the so-called dynamic puzzle is normally thought to apply specifically to self-deception, not necessarily to mere misrepresentation or self-delusion. I argued that the puzzle would apply directly to self-talk if understood, as I propose, as a communicative phenomenon. This was a reason to be less worried about the puzzle and, so, we should be less worried about it also in the case of self-deception. Self-delusion or self-misrepresentation, however, could not have served the same dialectical purpose, because they are normally not thought of as recruiting the same set of motivational or intention-like states.[7]

Robert Trivers and William von Hippel have argued for an interesting but controversial thesis about the evolution of self-deception (Trivers, 2011; Trivers, 2002, Chapter 8; von Hippel & Trivers, 2011). It is indeed puzzling, from an evolutionary standpoint, that humans evolved minds that systematically distort the truth by self-deception. Would it not be better, in the sense of increasing the individual's reproductive fitness more, to have access to one's own attitudes unsullied by such doxastic conflict? The solution to the puzzle, Trivers, and von Hippel argue, is to suppose that self-deception evolved in the service of other-deception. Many evolutionary psychologists agree that the ability to deceive others must have been extremely important in the evolution of human intelligence with clear positive effects on reproductive fitness, although some theorists argue that this is an exaggeration (e.g., Henrich, 2015). But, this ability presumably co-evolved with an ability to detect deception and, so, individuals must constantly try and discover new ways of getting away with their lies.

Trivers' idea is that deceiving oneself that, say, *p* is true makes it easier to deceive others that *p* is true and get away with it. And so there is a well understood evolutionary pressure on individuals to self-deceive under certain conditions. More specifically, conscious insincerity requires

[7]Thanks to an anonymous reviewer for this journal for pointing out that this needed to be made more explicit.

more cognitive effort than simply speaking truly, resulting in predictable and systematic behavioral manifestations. Increased cognitive effort makes people blink their eyes less, use longer pauses, and makes it more difficult to perform other tasks simultaneously (Trivers, 2011, p. 10). These behavioral cues are relatively easy to pick up on and they can thus give rise to suspicions that one is being lied to. And so, we should predict that many deceivers will adaptively and automatically seek strategies that minimize obvious cues of increased cognitive load. This is the evolutionary function, the argument goes, of self-deception. If I have deceived myself into thinking that *p*, I will not experience any increase in cognitive load when telling someone that *p*. My very belief that *p*, even if it is a form of self-deception, is formed in the service of making it easier for me to get others to believe that *p*, regardless of its actual truth value.

Funkhouser (2017) has recently claimed that this argument has some weak points, the chief one being that self-deception, when properly understood as distinct from self-delusion, does not in fact help deceivers decrease cognitive load. As we have seen, it is essential to self-deception that the deceived retain some awareness of the truth, stored in some part of the mind, even if only temporarily. If so, the source of the increase in cognitive load is still present when one tries to deceive someone that *p*, even if one is self-deceived that *p*, for that source is the deceiver's belief that *p* is not true. It seems then that only the delusional belief that *p* could afford one with the benefits of not manifesting cues of cognitive load when deceiving someone about *p*. But, as some theorists point out, we would normally not call this deception or lying: The person in question is simply saying what they consciously believe to be true (Fridland, 2011; Vrij, 2011). And, anyway, the costs of complete self-delusion would tend to outweigh the benefits of being better at deceiving others (Funkhouser, 2017, p. 226).

The objection, however, is inconclusive. On anyone's account of the mental state of being self-deceived about *p*, it partly consists in, or at least grounds, a disposition to self-represent as a *p*-believer. Minimally this means that being self-deceived that *p* involves assenting to *p* if the question of *p*'s truth comes up in one's inner reflections. Deep down one knows that *p* is false, one just carefully avoids reflectively assenting to *not p*. If this is so the self-deceiver always has a slight advantage over someone who simply believes that *not p* but wants to tell someone, insincerely, that *p*. The latter has no disposition in place to perform an act to self-represent as a *p*-believer. The former can, with less cognitive effort than otherwise, simply activate the self-directed disposition to signal being a *p*-believer whenever the issue comes up in the company of others. And so, it is indeed less likely that the self-deceiver lets off cues of cognitive load; there is a standing disposition to represent as a *p*-believer, even if one happens not to be a *p*-believer.

It does not follow, then, that only self-delusion could serve to decrease cognitive load. Further, von Hippel and Trivers provide an example to respond to the claim that we would not call this deception or lying:

> [I]magine I want to convince you that your spouse was not with my best friend while you were out of town. Imagine further that I have an acquaintance who mentions that he saw your spouse at 3:00 p.m. in the hair salon and at midnight in a bar. If I choose not to ask my acquaintance whom your spouse was with, or if I only ask my acquaintance whom she was with in the hair salon and avoid asking the more probative question of whom she was with in the bar, then I am lying when I later tell you that to the best of my knowledge she was not with my friend. Strictly speaking, what I am telling you is true. But the lie occurred when I initially gathered information in a biased manner that served my goal of convincing you of my friend's innocence regardless of what the truth might be. (von Hippel & Trivers, 2011, pp. 47–48)

Indeed, it seems implausible to insist that this does not count as an act of insincerity or lying merely because that is not how these words are ordinarily used. Perhaps it is better to change our use. Thus, we could opt for a theory of insincerity that cuts the connection, anyway based mostly on folk psychology, between insincerity and *conscious* attitudes. Dropping, at least, the idea that conscious intent is necessary for acts of insincerity, which is assumed in pressing this objection. But the sufficiency of conscious intent is not put into question by the example.

Even if these considerations fall short of vindicating the ambitious thesis that self-deception is a human trait selected for by natural selection because of a pressure to develop into more and more sophisticated other-deceivers, it seems safe to say that individuals might be motivated, partly or fully, to self-deceive in certain cases in order to be better at fooling others. This type of behavior could be adaptively learned, thus manifesting itself most distinctly in professions where persuasion is a prized commodity, politics being the most obvious candidate. As I understand his argument, Funkhouser would not disagree too strongly with this conclusion. He seems to think self-deception does not serve one overarching function, evolutionarily or otherwise, citing various possible functions instead. But, the bald man who deceives himself into thinking he is not bald could do so, at least in large part, because he wants others to believe he is not bald. Why should he care at all if he were socially isolated, only ever deceiving himself about something that normally matters merely insofar as it influences the reactions of others?

To conclude, consider the case of Peg and Greg again. Assume that Peg is self-deceived about her liking for Bach. In truth, she only performs the act of self-representing as a Bach-fan because she desires to be part of the prestigious Bach-club in Ballsbridge, where she lives. To become a member, she knows, it is very important that fellow Ballsbridgers believe that she is a genuine aficionado. Now, if she utters (1) in Greg's presence she could be accurately described as motivated by a desire to produce the belief that she likes Bach, in both Greg and herself. Moreover, she wants to produce this effect in both because, if successful, it becomes more likely that other Ballsbridgers come to have the same belief. If this is indeed possible, there is no *essential* difference in the motivational profiles of self-deceivers and other-deceivers even if, as a matter of statistical fact, people's reasons will fall into natural clusters depending on whether they are self-concerned or not in a particular case.

5 | CONCLUSION

In summary, my argument combines three basic ideas—ideas about representation, introspection, and self-deception—to reach the conclusion that other-talk and self-talk are alike in surprising ways. That is to say, they recruit the same mechanisms and share motivational and functional profiles.

1. Representational Acts are intentional actions whereby one represents oneself as having some attitude or being in some mental state. The competence to perform such acts is ultimately explained in terms of some cognitive mechanism whose function is to have specific effects on beings with the capacity to respond rationally to such acts. A plausible and popular hypothesis is that the acts are consumer-directed, that is, intentionally designed to cause predicted changes in the mental state of the recipient.
2. Introspection is a process whereby one represents oneself *to oneself* as having some attitude or being in some mental state. Moreover, the representational step which I have called Enactment is an intentional action directed at a rational agent with the aim of producing specific cognitive effects. Thus, if the competence to perform representational acts generally

is best explained by a Gricean mechanism, it is plausible that introspection recruits this very mechanism. This would mean that self-directed representational competence is not essentially different from other-directed competence.

3. SELF-DECEPTION, or at least a central aspect of self-deception, consists in a state in which the thinker's attitude contradicts their self-directed representational acts or dispositions to perform such acts. But even here, the acts in question could be other-directed or both self-directed and other-directed, while still being an integral part of the conflicted state in question. Moreover, the motivational profiles of deceptive representational acts do not vary in any essential respect across the divide between self and other. In particular, my primary aim in deceiving myself can be to deceive others and, it is reasonable to think, my primary aim in deceiving others can be to deceive myself.

Before going into the consequences for social epistemology, as promised in the title, I should make one comment on the mechanism shared by acts of speaker meaning and introspection. Strictly speaking, I have only shown that the two competences share a very basic structure, that is, the intentional production of a representational act to produce a cognitive state in a rational agent. The Gricean mechanism, however, is hypothesized to have a complex intentional structure dividing into, at least, the effective intention and the signaling intention.

I have argued, so far, that the effective intention is shared by the two mechanisms, but what about the signaling intention? This is both a more interesting and complex question than one might think, so I hope to address it in full on another occasion. Roughly, however, the signaling intention is present but even less effortful and further away from conscious accessibility. To see this, think again about the popular explanation for AVH. The idea is that inner speech is produced but the monitoring mechanism is flawed, so the thinker fails to self-attribute the production of the act. The failure cannot consist merely in not recognizing an effective intention, because the act is a full-blown act of speaker meaning and is definitely experienced as such in the hallucination (Wilkinson & Bell, 2016). An act of speaker meaning includes the signaling intention and, so, this is evidence for thinking that inner speech more generally includes that intention. Finally, inner speech is, as I have argued, merely one possible behavioral implementation of the representational act involved in introspection.

An important consequence to draw from the whole argument in this article is, simply, that every single phenomenon of interest to social epistemologists will apply directly, normally without any modification, to self-directed speech acts. Epistemologists tend to frame their discussions in terms of "testimony", a term I have avoided, but another way to put this is to say that testifying to oneself is subject to the same risks and biases as testifying to others. I will not apply this result in detail to specific cases here, but I assume that the way to proceed is fairly obvious. To take one brief example, consider the phenomenon of evidential preemption in Endre Begby (2021). Someone tells you *p* and adds that you should beware of any contrary evidence, suggesting that all such evidence has already been considered or will be misleading. As Begby puts it, this may "inoculate" the audience from contrary evidence and, possibly, be a form of epistemic manipulation.

The only point I want to make here is that we are at risk from self-preemption just as we are at risk from the possible manipulation of others. At some general level, this might seem obvious. Of course, one might think, this is *possible*. But if my argument is right things are in fact much worse. Basically, we have good reason to suppose that the cognitive mechanism in virtue of which evidential preemption works—namely the Gricean mechanism of speaker meaning—is also present in self-directed acts. That is to say, it is at work even in IPs, although many theorists have thought we can have some degree of immediate access to some of our own

attitudes and mental states. And if mechanisms and motivations are not essentially different in the two cases, we are at risk from self-induced evidential inoculation just as much as we are at risk from the manipulative preemptions of others. This realization, simple as it may appear at first, has very significant ramifications for social epistemology and related areas. The most immediate of these, perhaps, would be the eradication of the *interpersonal bias* of work in social epistemology. This work predominantly concerns itself with purported information exchange in personal *interactions*. But the effects and phenomena being scrutinized are already at work in muddying our access to our own first-person mental states.

ACKNOWLEDGEMENTS

Many thanks to Dan Harris for telling me that my long sermons on introspection should be turned into an article. Seems like he was right. I also thank James Norton and two anonymous referees for this journal for their helpful comments and suggestions. The ideas in this article grew naturally out of some arguments I presented in my book, *Talking about* (2022), especially Chapters 3 and 4.

ORCID

Elmar Unnsteinsson https://orcid.org/0000-0001-5333-1784

REFERENCES

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931–965.

Austin, J. L. (1975). *How to do things with words*. Clarendon Press.

Bach, K., & Harnish, R. (1979). *Linguistic communication and speech acts*. MIT Press.

Bar-On, D. (2004). *Speaking my mind: Expression and self-knowledge*. Oxford University Press.

Bechtel, W. (2007). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Begby, E. (2021). Evidential preemption. *Philosophy and Phenomenological Research*, *102*(3), 515–530.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford University Press.

Carruthers, P. (2018). The causes and contents of inner speech. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 31–52). Oxford University Press.

Carston, R. (2002). *Thoughts and utterances*. Blackwell.

Chomsky, N. (1975). *Reflections on language*. Pantheon Books.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Clarendon Press.

Fernyhough, C. (2016). *The voices within: The history and science of how we talk to ourselves*. Basic Books.

Fridland, E. (2011). Reviewing the logic of self-deception. *Behavioral and Brain Sciences*, *34*(1), 22–23.

Funkhouser, E. (2005). Do the self-deceived get what they want? *Pacific Philosophical Quarterly*, *86*(3), 295–312.

Funkhouser, E. (2017). Is self-deception an effective non-cooperative strategy? *Biology and Philosophy*, *32*(2), 221–242.

Funkhouser, E., & Barrett, D. (2016). Robust, unconscious self-deception: Strategic and flexible. *Philosophical Psychology*, *29*(5), 1–15.

Garson, J. (2013). The functional sense of mechanism. *Philosophy of Science*, *80*(3), 317–333.

Garson, J. (2017). A generalized selected effects theory of function. *Philosophy of Science*, *84*(3), 523–543.

Garson, J. (2022). Putting history back into mechanisms. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/715112.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Gregory, D. (2016). Inner speech, imagined speech, and auditory verbal hallucinations. *Review of Philosophy and Psychology*, *7*(3), 653–673.

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

Grice, P. (2001). *Aspects of reason*. Oxford University Press.

Harris, D. W. (2021). Semantics without semantic content. *Mind & Language*, *37*(3), 304–328.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Henrich, J. (2015). *The secret of our success*. Princeton University Press.
Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.
Langland-Hassan, P., & Vicente, A. (2018). *Inner speech: New voices*. Oxford University Press.
Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.
Lycan, W. G. (1996). *Consciousness and experience*. MIT Press.
Mele, A. R. (2001). *Self-deception unmasked*. Princeton University Press.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
Moran, R. A. (2001). *Authority and estrangement: An essay on self-knowledge*. Princeton University Press.
Neale, S. (1992). Paul Grice and the philosophy of language. *Linguistics and Philosophy*, *15*(5), 509–559.
Neale, S. (2005). Pragmatism and binding. In Z. G. Szabó (Ed.), *Semantics versus pragmatics* (pp. 165–285). Clarendon Press.
Neale, S. (2016). Silent reference. In G. Ostertag (Ed.), *Meanings and other things: Themes from the work of Stephen Schiffer* (pp. 229–344). Oxford University Press.
Neander, K. (2017). *A mark of the mental: A defence of informational teleosemantics*. MIT Press.
Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*(2), 97–132.
Rosenthal, D. M. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 109–131). Oxford University Press.
Schiffer, S. (1972). *Meaning*. Oxford University Press.
Schiffer, S. (1987). *Remnants of meaning*. Cambridge University Press.
Schiffer, S. (2003). *The things we mean*. Clarendon Press.
Schwitzgebel, E. (2019). Introspection. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University.
Scott-Phillips, T. (2015). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Palgrave Macmillan.
Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
Simons, M. (2017a). Local pragmatics in a Gricean framework. *Inquiry: An Interdisciplinary Journal of Philosophy*, *60*(5), 466–492.
Simons, M. (2017b). Local pragmatics in a Gricean framework, revisited: Response to three commentaries. *Inquiry: An Interdisciplinary Journal of Philosophy*, *60*(5), 539–568.
Sperber, D. (2013). Speakers are honest because hearers are vigilant: Reply to Kourken Michaelian. *Episteme*, *10*(1), 61–71.
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.
Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Blackwell.
Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell.
Trivers, R. (2002). *Natural selection and social theory: Selected papers of Robert Trivers*. Oxford University Press.
Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others*. Penguin Books.
Unnsteinsson, E. (2022). *Talking about: An intentionalist theory of reference*. Oxford University Press.
Vicente, A., & Jorba, M. (2017). The linguistic determination of conscious thought contents. *Noûs*, *53*(3), 737–759.
von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1–16.
Vrij, A. (2011). Self-deception, lying, and the ability to deceive. *Behavioral and Brain Sciences*, *34*(1), 40–41.
Wilkinson, S., & Bell, V. (2016). The representation of agents in auditory verbal hallucinations. *Mind & Language*, *31*(1), 104–126.
Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.
Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Harvard University Press.

How to cite this article: Unnsteinsson, E. (2023). The social epistemology of introspection. *Mind & Language*, *38*(3), 925–942. https://doi.org/10.1111/mila.12438