Author's copy of a paper published in *The Philosophical Quarterly* in 2021.

Being self-deceived about one's own mental state

By Kevin Lynch

Abstract

A familiar puzzle about self-deception concerns how self-deception is possible in light of the paradoxes generated by a plausible way of defining it. A less familiar puzzle concerns how a certain type of self-deception–being self-deceived about one's own intentional mental state–is possible in light of a plausible way of understanding the nature of self-knowledge. According to this understanding, we ordinarily do not infer our mental states from evidence, but then it's puzzling how this sort of self-deception could occur given that self-deception arises from the mistreatment of evidence. This article argues that to accommodate this kind of self-deception we should accept that sometimes ordinary self-knowledge is inferential, but that this idea needn't be so unappealing. In particular, by showing that such inferential self-knowledge can be both 'transparent' and 'direct', the article argues that it need not imply having an abnormal, 'alienated' relation to the mental state.

> It makes sense to ask, 'Do I really love her, or am I only fooling myself?', and the process of introspection is the calling up of memories, of imagined possible situations, and of the feelings that one would have if ... (Wittgenstein, *Philosophical Investigations,* §587)

1. A tension between three views

A distinction can be made between two kinds of self-deception, one which we may call 'world-directed self-deception' (WDSD) and the other 'mind-directed self-deception' (MDSD). In WDSD the subject deceives himself into believing something about a matter outside his own mind, and in MDSD the subject deceives himself into believing something about his own mind, about whether or not he has a

certain mental state.[1] To illustrate, consider where one person in a marriage has fallen out of love with the other, say the wife with the husband. It might be difficult for both of them to believe this is true. Both might persist in believing that she still loves him, despite numerous signs to the contrary. Further, both might be self-deceived in believing this, but for the husband this would be WDSD while for the wife it would be MDSD. The matter of the wife's feelings for her husband is a matter external to the husband's mind, but not the wife's. The wife is deficient in *self*-knowledge, but not the husband.

This article is about MDSD, about being self-deceived about our own *intentional* mental states specifically (e.g. beliefs, desires, aversions, and emotions). MDSD has rarely been treated as a special topic in itself, which is unfortunate since there are philosophical problems that are peculiar to it. For it seems puzzling, given certain natural assumptions, that MDSD should even be possible, except in special circumstances. This is not for the familiar reasons that seem to militate against the possibility of self-deception of any sort (namely, the paradoxes that a certain way of defining 'self-deception' gives rise to), but for reasons related to the alleged nature of self-knowledge, reasons that apply only to that kind of self-deception.

The problem I have in mind here is manifested in a tension that exists between three popular and plausible views. *View 1* is the claim that MDSD happens. This is widely taken for granted, as one can gather from the many occasions in the self-knowledge literature where MDSD is adduced to discredit Cartesian infalliblism about our own mental states. However, I take it that our assumption here is not simply that MDSD exists, but that it is something we are susceptible to *in the normal course of events*. It does not seem as though this can only happen in abnormal circumstances, such as when the mental state is *inaccessible* or *unconscious* due to 'repression' or cognitive malfunction. For instance, we needn't think that Wittgenstein's man, in the above epigraph, is uncertain about his feelings for a woman because those feelings are inaccessible (more on this shortly).

View 2, which is more moot than the others but still very popular, is that when we do get things right about our own mental states in this 'normal course of events', our self-knowledge is 'immediate', which is to say non-inferential or *not based on evidence*. On this view, we do not wait to see how we behave or what we say in order to know what we feel or think about something, and neither do we infer our mental states from internal goings-on: inner speech, imaginings and the like. Knowledge of our

[1] I am assuming a doxastic view of self-deception here, according to which self-deception involves having a false belief held against good evidence to the contrary. This is the most popular view of self-deception and I have defended such a view in other articles (e.g. Lynch 2017; 2012).

own minds on this view is markedly different from knowledge of other minds, where evidence is needed. Subscribers to View 2 do think that we can know of our own mental states on the basis of evidence, but they regard these sorts of cases as abnormal, some even say 'pathological' (Boyle 2019: 1015).[2] Let us call the view that in normal cases of self-knowledge the knowledge is not based on evidence, 'the immediacy assumption'.

View 3 concerns the etiology of self-deceptive beliefs. When we observe accounts of the processes that lead to self-deceptive beliefs, we invariably see that they describe some kind of *mistreatment of evidence*. This could include being hypercritical towards or rationalizing unwelcome evidence, ignoring evidence, and selectively gathering or focusing on evidence. In other words, self-deceptive beliefs come about through one's dealings with evidence, and therefore in cases where one's epistemic access to the relevant fact is evidence-mediated. It is difficult to imagine what a self-deceptive process could otherwise be.

These three views are incongruous because, apparently, if self-deception can only occur when our epistemic access to the relevant fact is evidence-mediated then the possibility of MDSD occurring in the normal course of events is excluded, since that is when we are supposed to not require evidence at all to know our own minds. That is to say, MDSD only seems possible when we are reliant on evidence to know our own minds, but such cases are supposed to be abnormal or deviant, where the mental state is 'inaccessible to direct introspection'. Surely one of these views must yield to the others. But which one?

The aim of this paper is to resolve this tension, and in the process to clarify how MDSD is possible. In section 2 we will examine a previous attempt to do the same which effectively rejects View 1, holding that MDSD occurs only with mental states that are *unconscious* (and everyone agrees we use evidence to know those). Against this, I contend that paradigmatic MDSD involves mental states that are accessible to introspection. Assuming View 3 this would suggest that, at least sometimes, ordinary self-knowledge comes about through inference from evidence, a thesis we may call *inferentialism*. However, as section 3 explains, there is a common view in the literature–especially heard from philosophers who espouse the so-called 'rational agency' theory of self-knowledge–that inferential self-knowledge is by its very nature of an abnormal and 'alienated' kind, an idea I trace to the assumptions that, unlike normal self-knowledge, inferential self-knowledge lacks 'transparency' and is

[2] There are some philosophers (e.g. Carruthers 2011; Cassam 2014) who argue that ordinary cases of self-knowledge are inferential, but they portray themselves as battling orthodoxy.

'indirect'. In response to this, in section 4 I develop a model of mental states that explains how self-knowledge can be inferential, and then in sections 5 and 6 respectively I argue that such inferential self-knowledge can be transparent and direct. Thus the paper resolves the tension by rejecting View 2 and arguing that this isn't as unappealing as one might think. This opens the way to understanding how MDSD can happen in the normal course of events. In the final section some specific psychological processes are discussed to help show more concretely how MDSD is possible.

In the attempt to clarify MDSD and develop an explanatory model for it, this paper will rely on the case of romantic love, and to a lesser extent, desire. This may provoke a criticism, in that it could be questioned how far the model generalizes to other mental states, especially belief, which the mentioned rational agency theorists have mostly focused on. Let me state that it is not the ambition of this paper to provide a model for explaining all MDSD (though I will have some things to say about belief). But as we will see, the alienation objection to inferential self-knowledge advanced by these theorists is one that is applicable to a broad range of mental states, apparently any mental state that is intentional and reasons-responsive. If correct, it could rule out the possibility of MDSD occurring in the normal course of events for that whole class. Thus it would be a significant result if we could show a way of meeting this objection for at least some of these mental states.

2. A previous attempt to solve this problem

In a recent article, Mathieu Doucet noted the above tension. Doucet is concerned to defend the 'rational agency model' of self-knowledge of our own 'first-order intentional states', and he sees MDSD as posing a threat to it. The details of the rational agency model are not important here. What is important is just that it is committed to View 2, the immediacy assumption (2011: e3). The problem, then, is that if we assume that self-deception happens due to a biased treatment of evidence (2011: e8), then '[s]ince the rational agency model denies that self-knowledge is formed on the basis of evidence, this would seem to rule out the possibility of self-deception with regard to … our intentional states' (2011: e3), an unwelcome implication in Doucet's opinion. Note that a number of other theories hold that such self-knowledge is typically not based on evidence (e.g. expressivist, constitutivist and perhaps also introspectionist views), so this problem is not peculiar to the rational agency model.

Doucet tries to resolve this tension basically by targeting View 1. He does accept that MDSD can happen, but he believes that this doesn't threaten the idea that self-knowledge is *normally*

immediate. This is because cases where we are reliant on evidence to know our own minds, and where MDSD is thus liable to occur, are *ab*normal: they are cases where something has 'gone awry with [our] self-understanding' (2011: e4), where something that should be immediately knowable becomes only mediately knowable through the intermediary of evidence. Doucet describes subjects as being *estranged* or *alienated* from their mental state in such circumstances. Accordingly, his model example of MDSD is a case of a man undergoing psychoanalytic therapy. His therapist studies his behaviour, free-associations and so on, and she arrives at an interpretation, telling him that he believes he is unworthy of success. This comes as news to him, but as welcome news since it 'flatters him by providing a convenient explanation for his lack of success that he prefers to the alternative, which is that he simply lacks the talent required to succeed' (2011: e16). Having heard what he wants to hear, the patient uncritically accepts the interpretation despite having good reasons to doubt it: 'perhaps the conclusion is out of keeping with all of his other beliefs, or his therapist has proved to be mistaken in the past, or his therapist's examination was cursory and distracted' (2011: e16-e17). He is thus self-deceived in believing that he has this belief.

Although Doucet succeeds in making room for the possibility of MDSD, his approach is not entirely satisfactory since it doesn't seem that MDSD only happens with mental states that are unconscious, as Doucet thinks (2011: e23). As can be seen from some literary examples, paradigm cases of MDSD look quite different from Doucet's therapy case and do not involve mental states that are inaccessible to introspection, at least not in any deep, psychoanalytic sense. Consider this description from Tolstoy's Anna Karenina of the eponymous protagonist:

> There she met Vronsky and experienced a disturbing joy at these meetings … Vronsky was everywhere he might meet Anna, and he spoke to her whenever he could of his love. She gave him no encouragement, but each time she and he met, her soul burned with the same feeling of animation that had descended upon her that day in the train car when she had seen him for the first time. She herself could feel the delight shine in her eyes at the sight of him and her lips furrow into a smile, and she could not suppress the expression of this delight.
>
> At first Anna sincerely believed that she was displeased with him for allowing himself to pursue her; however, soon after her return from Moscow, upon arriving at a party where she thought she might but did not meet him, she distinctly realized from the disappointment that

came over her that she had been deceiving herself, that not only did she not find this pursuit unpleasant, but it constituted the entire interest of her life (Tolstoy 2014[1878]: 118-119).[3]

Notice how Tolstoy describes Anna as *feeling* the delight, of her soul *burning* in Vronsky's company, and as being *unable to suppress* her joy: she was assailed by conscious feelings that were upsurges of her love for Vronsky. Clearly, then, her love for him was not inaccessible. The feelings associated with love were experienced but they were *misinterpreted* by her, attributed to something other than love. (It is a fictional case of course, but it does not seem far-fetched.) So it seems that there is something about the nature of an intentional mental state like love that makes MDSD possible with it even when it is accessible to introspection. Surely these will be the most prevalent and interesting cases of MDSD, and ones we should strive to understand. But this seems to require accepting an inferential element at the heart of much ordinary self-knowledge. However, if this is to be our position, then the worries that many philosophers have about inferential theories of self-knowledge must be addressed. Let us examine next a key worry that's frequently voiced in the literature.

3. Inference and alienation

We've seen that MDSD can occur in ordinary circumstances when the relevant mental state is introspectively accessible. This would imply that in at least some ordinary cases of self-knowledge we rely on evidence to know our own minds, a view I'll call inferentialism about self-knowledge.[4] However, many philosophers have had serious misgivings about such a view. One of the deepest of these is the idea that such reliance on evidence would imply being in the abnormal condition mentioned by Doucet: being 'estranged' or 'alienated' from the mental state (the *locus classicus* of this idea is in (Moran 2001), especially §3.3; also see (Boyle 2015)).

I believe there are two strands to this idea. One will be discussed in section 6 but the other, as I understand it, goes as follows. Inferentialists about self-knowledge hold that there is no deep difference

[3] A similar case is in Marcel Proust's *Remembrance of Things Past* concerning the relationship between the narrator and Mademoiselle Albertine. See the beginning of volume 6 in particular. This case is discussed in (Nussbaum 1988).

[4] Thus defined, inferentialism can come in moderate and hardline varieties, depending on how much ordinary self-knowledge we see as inferential. So for instance, Krista Lawlor would be among the moderates while Quassim Cassam and Peter Carruthers–who think that all self-knowledge within certain classes are inferential–are among the hardliners. I only wish to defend a moderate inferentialism here and I have expressed skepticism of hardline inferentialism elsewhere (Lynch 2016).

between how we know our own minds and the minds of others. As the original inferentialist Gilbert Ryle put it, 'John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe' (2009[1949]: 138). But consider now the latter kind of case. When I observe another person S's actions or reactions, or listen to what he says, and form a belief on that basis that, say, he believes that P, this might not incline me in the least to believe that P myself (unless *per accidens* I have certain auxiliary beliefs such as that S is knowledgeable on this matter). The evidence that S believes that P is not evidence that P, and I might be aware of no reasons to believe that P, in fact, I might even have good reasons to believe that not-P. Now if we know our own minds in a similar manner then we are entitled to wonder why things would be any different. That is, if I discover what I believe by drawing conclusions from similar pieces of evidence exhibited by myself, then why would I treat 'my belief' any differently than I would the belief of another? (Why should I be inclined to use it as a premise in reasoning, say, or act on the assumption that it is true?) For the pieces of evidence indicating that I believe that P *are not pieces of evidence indicating that P*, and my being aware of them is compatible with my being aware of *no* evidence that supports P. Importantly, the same applies to other intentional mental states: evidence that S fears X is not necessarily evidence that X is threatening or dangerous, evidence that S loves X is not necessarily evidence that X is lovely, or evidence that S wants X is not necessarily evidence that X is desirable or good, whether S happens to be another person or oneself. Here we see the appropriateness of talk about being alienated from the mental state. If my own mental state becomes known to me through inference from evidence, then why should I *identify with* or *endorse* it any more than I would the mental state of another person that I detect in a similarly inferential way? For one's awareness of the mental state is 'detached from [one's] sense of the reasons … supporting it' (Moran 2001: 93).

The inferentialist Quassim Cassam (2014: chap. 11) dismisses this objection by saying that there is no incompatibility between knowing your own belief inferentially and endorsing it. But Moran would grant this, just as he would accept that one can infer that another person believes that P and also endorse that belief. The point is that the act of inferring the belief doesn't cause the endorsement. Endorsement would come from a *separate act* of apprehending evidence/reasons for P. The connection between the self-knowledge and the endorsement would be accidental.

For these critics of inferentialism, alienated self-knowledge is supposed to contrast with 'transparent' self-knowledge: self-knowledge that results from a 'transparency procedure'. In such cases, 'when a subject is asked whether she has a certain propositional attitude, she addresses the

question by focusing her attention on the intentional object of the relevant attitude' (Fernández 2013: 18). For instance, if S is asked whether she wants to V, in answering this she will consider V-ing, weighing the pros and cons of doing that. Finding only pros she might then declare, 'That would be great,' or 'I'd love to'. But she is not thinking *about herself* in this process: her attention is directed 'outward', at the activity of V-ing and what it would involve. (Perhaps wanting to V is not a *propositional* attitude, but this would just show that the transparency procedure is not limited to the propositional attitudes.)

I do not think that we should regard this as a procedure for acquiring self-knowledge as such, certainly not knowledge of a pre-existing attitude (a point now widely accepted). It is rather a procedure for *forming* an attitude, for making up one's mind about something, which may bring self-knowledge of that attitude in its wake. (Analogously, flipping the light-switch is not the procedure for *knowing* that you have turned on the light, rather, it's the procedure for turning on the light, though this typically results in you knowing that you have turned on the light.) In this I am in agreement with a number of transparency theorists, who hold that more is required for self-knowledge than a mere outwardly-directed judgment, though opinions diverge on what else exactly is needed (more on this issue shortly).

Nevertheless, we can still appreciate how any self-knowledge born from such a procedure would not be alienated, since it will not be 'detached from [one's] sense of the reasons' for the attitude. It will be self-knowledge that comes from considering the intentional object. Inferential self-knowledge, on the other hand, needn't be associated with considering the intentional object of the attitude at all, which is why alienation would result. Doucet's man, for instance, needn't have considered whether he is unworthy of success before he heard and accepted his therapist's assertion that he believes that.

I take this problem of alienation to be a serious one for an inferentialist theory. However, I will suggest that the problem lies not with the very idea of self-knowledge gained by inference as such. Rather, whether inferential self-knowledge is of an alienated kind would depend on the kinds of evidential sources we infer from. We are held back here, I believe, by a false dichotomy. This dichotomy is between self-knowledge that is, on the one hand, immediate or direct, non-inferential, 'transparent' and 'normal', and oppositely, mediate or indirect, inferential, not transparent, 'alienated' and 'abnormal'. But an intermediate type of case can be recognized: self-knowledge that is inferential while also being transparent and, in an important sense, direct. Realizing this will make an inferential

theory of self-knowledge much more palatable, which in turn will open up some logical space for MDSD.

4. External and internal evidence

Recent inferentialists have distinguished between external and internal evidence for inferring our own mental states. External evidence includes our own behaviour and the circumstances we are in. Internal evidence includes things like thoughts, feelings, imaginings and rememberings. Though older inferentialist theories emphasized the former (Ryle 2009[1949]; Bem 1972), more sophisticated recent versions emphasize the latter (Cassam 2014; Lawlor 2009; 2008).

At this point it will be useful to have some examples to hand to see what these evidential sources, and the self-knowledge based on them, can look like. Let's follow the lead of Wittgenstein and Tolstoy and continue with the case of romantic love, which, though not obviously a *propositional* attitude, is like belief, desire and fear in being an intentional and reasons-responsive mental state. So suppose that a man, call him Romeo, loves a woman, Juliet. The following, then, are some ways that this love could psychologically manifest itself. They are, arguably, universal features of love, and things that could serve as the evidential basis for Romeo self-attributing that state.

1) Juliet occupies Romeo's thoughts a lot. She is often on his mind.
2) He feels happy when he is with her.
3) When he is not with her for an extended period, Romeo misses her. This includes feeling a longing to be reunited with her, and feeling excited about their reunion.
4) When Romeo is reunited with Juliet after an extended period, he feels delight.
5) If she rejected him or left him he would feel deeply and persistently upset.
6) If another man wins her heart he will feel strong disappointment and will experience feelings of jealousy, hostility and anger.
7) Romeo cares about Juliet and his own moods, thoughts and emotions fluctuate along with her fortunes, for example, he will fret if she is in trouble, and feel delight at her successes.
8) He sees her as special compared to other potential partners. He has lost interest in other women, or that interest has lessened considerably.

This is only meant as a minimal characterization of just some aspects of love, and much more could be said about them (for instance, if X loves Y then yes, Y is often on X's mind, but he/she must be on it in a particular way). A number of general observations can be made about this mental state, based on the above, and we could expect these points to apply to some other mental states to some degree or another.

Structural complexity: Though we often speak of love as 'a feeling' this is incorrect; it is not just *one* sort of feeling, but many. Love will manifest in different feelings–of joy or sadness, satisfaction, longing, excitement, disappointment, contentment, emptiness (after bereavement), worry, anger or envy–depending on the circumstances, depending on whether one's love is requited or unrequited, gained, lost, taken from one, in trouble, and so on (and there is no good reason to single out any one of them as being what love essentially is). This might make us feel uneasy about calling love a 'mental state', which connotes a constancy that it does not have (unlike paradigmatic mental states like a depressive mood), though I will follow tradition and continue to call it that.

Suppressability: The persistent absence of these manifestations would be strong grounds for denying that Romeo loves Juliet, but it's nevertheless possible for them not to manifest while the attribution applies. But then there should be a *special explanation* for this lack. For instance, maybe Romeo didn't think of her much on one day, but only because he was extremely busy in work or caught up in some crisis. In other words, such manifestations can be absent in their usual circumstances but only because some factor is suppressing them.

Logical (or rational) connectedness: These different manifestations are not some contingent assemblage, the way the different properties or parts of some physical objects are, but constitute a rationally connected whole. It is not, for example, a contingent fact that when you miss another person you tend to feel delight, rather than disappointment say, at meeting them again, the way it is a contingent fact that oranges are round. For part of what it means to miss someone is that you want to be reunited with them (though you might not want this all-things-considered), and it's pleasing to get things that you want. It would be irrational to miss someone and then be displeased in his/her company (unless some countervailing factor intervenes).

This model for understanding certain mental states as structurally complex, consisting of *logically-related suppressible manifestations* (which in section 6 I'll argue are genuine parts of the mental state, rather than effects of it), helps to make sense of how self-knowledge of these mental states can require inference. This is because such manifestations can serve as *internal signs or evidence* for the associated mental state. For instance, Romeo might notice how much he misses Juliet when she is away, or how much she has been on his mind recently, or the pang of jealousy or hostility he feels when he sees her flirting with another man, and *conclude* on that basis that he loves her. Indeed, Tolstoy supplied a similar example above where Anna Karenina inferred 'from the disappointment that came over her' that she loves Vronsky. Inference is required here because these experiences are not what love is in its entirety but are manifestations of it. An inference must be made from, say, 'I miss her' or 'I'm so concerned about her' to 'I love her'. The latter claim then carries the implication that the other manifestations of love would occur given the right circumstances.

Desire is another mental state that fits this model well, since it too can manifest in different feelings, such as longing, anticipation and excitement, frustration, disappointment, envy, delight and relief. Krista Lawlor (2009) has illustrated how knowledge of one's own desires can involve inference, with her case of a woman, Katherine, who infers that she wants another child from inner, psychological evidence, or what Lawlor calls 'internal promptings'. More specifically, when 'putting away her son's now-too-small clothes,' Katherine 'finds herself lingering over the memory of how a newborn feels in one's arms [wistfully perhaps],' and on another occasion she experiences 'envy when an acquaintance reveals her pregnancy' (2009: 57). She concludes on that basis and more that she wants another child.

However, it would be a mistake to extend this model to all mental states, and to use it to account for all MDSD. For some mental states are not so structurally complex, like disappointment or joy (though MDSD might be possible for them). The mental states that serve as the internal evidence for the above inferences will commonly be of this sort. But how about belief? Belief, a non-affective mental state, is not associated with having particular *feelings* in specifiable circumstances the way love and desire are. Nevertheless dispositional profiles of a different kind can be associated with belief, for instance, believing that P can imply a willingness to act and take risks on the assumption that P, or to be critical of evidence for not-P. Other dispositions can depend on the content of the belief and on what other mental states one has (e.g. if you believe you were insulted you'll feel aggrieved). Perhaps such dispositional manifestations could be evidence for our own beliefs (see Lawlor 2008), evidence that could be mistreated in MDSD. But due to its differences from our focal mental states I cannot give a

satisfactory explanation of MDSD about belief in this article, though I'll have some things to say about it in the final section.

In this section we have seen how internal, psychological evidence can be used in inferential self-knowledge. The recent emphasis on internal evidence by inferentialists is a positive development. However, these theorists treat this evidence as just more evidence, just more data to support a self-attribution on a par with behavioural evidence. They fail to note some distinctive features of this internal evidence. Next I will argue, in the following two sections respectively, that inferring one's mental state from such internal evidence is compatible with the self-knowledge being both transparent and direct, and is therefore compatible with its normalcy.

5. Inferential and transparent

Transparent self-knowledge is self-knowledge associated with the 'transparency procedure'. This procedure involves 'looking outward', attending to or considering the intentional object of the known mental state, as opposed to oneself, and self-knowledge is supposed to result from this (though how exactly it does so is debated).

But if that's what transparent self-knowledge is, then it's unclear why the cases of inferential self-knowledge mentioned above should stand in contrast to it. For notice that some of the evidence that Romeo may use to know that he loves Juliet involves such 'looking outwards'. For instance, having her on his mind a lot, missing her, or seeing her as special all involve *attending to her*. They are thoughts and feelings *about Juliet*, directed at her 'loveable qualities', qualities that, for Romeo, rationalize his love. He notices that he has been thinking about her in a particular manner, a manner indicative of love, and then judges that he loves her.

Similar things can be said about Katherine. Consider her 'lingering over the memory of how a newborn feels in one's arms', which she takes as evidence that she wants another child. But this piece of evidence is a state of considering the intentional object of that desire: the prospect of having another child (the memory of her last child represents what she could expect from having another). Katherine considers that prospect and finds something appealing about it. She does something that she might do if she asked herself, 'Should I have another child?' or 'Would it be good to have another child?' She notices herself doing this and then concludes that she wants another child, but by the same token she is aware of a reason *to* want another child. Why would this not be a case of transparent self-knowledge?

The general reason why this inferential self-knowledge is also transparent is not difficult to see. The evidence in question here are thoughts, imaginings, memories and feelings. But these are intentional mental states too, states that are 'directed at an object' or that 'refer to a content'. So we are inferring one intentional mental state *from another* (a standing one from an occurrent one perhaps). It shouldn't be surprising, then, that this is transparent self-knowledge, if that means self-knowledge that comes from 'focusing … on the intentional object of the relevant attitude'. It also shouldn't be surprising that we can infer a standing intentional mental state from an occurrent state (or event), since many philosophers have held that standing states partly consist of dispositions to think and feel certain things on appropriate occasions.

An objection to the idea that these are cases of transparent self-knowledge would be made by Wolfgang Barz however. Though this objection is, I believe, misguided, it is an interesting one and forces us to think more clearly about what transparent self-knowledge is. I described Romeo and Katherine as *noticing themselves* thinking/feeling something and inferring from that, but for Barz this would rule them out as transparent self-knowers, simply because it attributes to them a self-directed thought or shift of attention towards themselves. This, according to Barz, is introspection; genuine transparency involves not introspection but 'extrospection': focusing 'exclusively' on the intentional object (2019: 918 & 924).

It was suggested above, however, that the transparency procedure by itself is just a procedure for forming an attitude and not for knowing one, and it's doubtful that a *mere* outwardly directed thought could be sufficient for self-knowledge (see Boyle 2019). The most we could say about these 'extrospectionist' cases is that the subject has a kind of *tacit* knowledge, since the proposition purportedly known to be true–that one has such-and-such attitude–*has not been considered* and the relevant concepts have not been applied, though the subject is poised to assent to it (thus they are akin to your knowing that the Earth's core isn't made of marmalade, which you knew and were poised to assent to before reading this sentence).[5] Barz's understanding of the transparency idea is extreme, and is born of the assumption of a sharp dichotomy between 'introspection' and 'extrospection'. The solution here is to note that attending outward and inward are not mutually exclusive. For one can have

[5] On finding that the evidence supports P one might say 'I believe that P', but this does not necessarily express self-knowledge. 'I believe that P', for instance, can be just a hedged or qualified assertion that P and needn't express a self-directed thought.

a certain regard for something in the world and then take note of the fact that one is regarding it so. It is here that inferential yet transparent self-knowledge can be found.

This section has argued that correctly inferring our mental states from internal evidence is compatible with that self-knowledge being transparent. But this isn't to say that all psychological evidence would result in transparent, unalienated self-knowledge, were we to use it for our inferences. Consider inner speech, which Ryle and Carruthers have said we rely on to know what we think. Now just because you hear *another* person voice an attitude, saying 'P' or 'I love so-and-so', does not mean that you will identify with that attitude, for you might never have given much thought to whether P or to so-and-so yourself. Similarly, just because you 'hear' yourself say 'P' or 'I love so-and-so' in inner speech doesn't necessarily mean you will identify with that expressed attitude either, for the very same reasons. So self-knowledge based on 'overhear[ing]' or 'eavesdrop[ping] on' (Ryle 2009[1949]: 165) our own inner speech would likely be of an alienated kind, but this is explicable in terms of inner speech not being a transparent source of evidence. 'Listening' to one's own inner speech involves attending merely to oneself, to what one is doing, and need not involve reflecting on what that speech is about.

6. Inferential and direct

There is another strand to the idea that inferential self-knowledge would involve alienation, based on a sense of alienation distinct from, though perhaps related to, the previously described sense. This previous sense was related to the idea of lacking a 'sense of the reasons' for the attitude, but the further sense is related to the idea of mediacy or indirectness. For some have suggested that inferential self-knowledge is by its very nature of an indirect sort, necessary only when we lack 'direct access' to the mental state. It thus implies being in a state of separation from the mental state which the pejorative term 'alienation' seems apt to describe, especially considering the numerous intellectual traditions that regard having direct access to our mental states as normal, healthy, optimal or ideal.

However, we must first distinguish two different concepts of epistemic mediacy, since mediacy can be defined in terms of facts or things (including events). The first more common concept is being expressed when it is defined as knowledge that 'depends for its status as knowledge on other knowledge' (Alston 1983: 73), or as knowledge that's '"based" or "grounded" or "founded" upon my knowledge of [some] other fact' (Moore 1929: 74). But there is another idea of epistemic mediacy–one

that's more important for understanding talk of 'access'–where we say that our knowledge *about some thing* is mediate because it is based on our knowledge of some *other thing* (e.g. inferring that x is F from seeing that y is G). This second concept can be found in, for instance, the indirect realist theory of perception, which holds that our awareness of physical objects is 'indirect' in being based on our awareness of intermediate entities: sense-data. Following this precedent, let's reserve the terms 'mediate'/'immediate' for the former concept and 'direct'/'indirect' for the latter. We will soon see that these distinctions cut across each other, and that though inferential self-knowledge is, by definition, mediate, it needn't be indirect and so needn't be alienated in the relevant sense.

6.1 Causal inference

The idea that inferential self-knowledge is indirect in the sense just stipulated is suggested by how inferentialist philosophers have conceptualized the inferential process. According to this conception, when we infer a mental state from some evidence we are inferring from *an effect to its cause*, which–given the ontological distinctness of cause and effect–entails that the self-knowledge is indirect. Again, other-person cases, where we gain knowledge of other minds, are treated as the model. It is commonly believed that when we perceive another person's behaviours, reactions or utterances we are witnessing the *effects* of mental states 'within' the person. Attributing beliefs, desires, and emotions to others is thus making an inference from these outward effects to their inner mental causes. This way of thinking then gets transferred by inferentialists to the first-person case. Knowing our own minds is then seen as a matter of inferring from various effects an inner mental cause, with the only difference being that the self-knower is privy to a wider range of effects, in particular the inner manifestations (with the inferred cause being more deeply inner still).

This way of thinking is most explicit in Lawlor's inferentialism, which has influenced that of Cassam. Lawlor philosophizes that 'causal self-interpretation' plays a significant role in much 'quotidian' self-knowledge, and she elaborates on how it occurs for belief (2008) and desire (2009). Such self-knowledge, she maintains, often involves noticing behaviours or behavioural patterns and also the mentioned 'internal promptings', and then 'making an inference about their likely causes' (2009: 49), which is a matter of positing intentional mental states that would explain them. Annalisa Coliva agrees that sometimes we infer the existence of mental states from our 'inner phenomenology', states that we posit as 'its probable causal explanation' (2016: 70). Though Peter Carruthers talks about interpretation rather than inference, his view of self-knowledge looks similar. For Carruthers, we know

our own propositional attitudes only by interpreting 'sensory cues' (2011: 69) such as sensations, imagery, inner speech and 'perceptual representations' of our own behaviour. These cues are the products of these propositional attitudes (53-54), attitudes that we lack 'direct access' to (69).

It is this causal conception of the inferences involved in inferential self-knowledge, I suggest, that further creates a problematic picture of having an alienated relation to these intentional mental states. Causes and their effects are supposed to be distinct from each other. On the view being considered we are aware, perhaps in a direct way, of the effects but we hypothesize or posit the cause and are thus at some remove from it.[6] These hypotheses are merely 'probable' or 'likely'. The phenomenology of intentional mental states is portrayed as a phenomenology of mere symptoms: the mental state itself is something that 'underlies' (Lawlor 2009: 73) that symptomatology. It is an I-don't-know-what that produces these symptoms and is not a part of our stream of consciousness. We are thus alienated from it in something like the way that we are alienated from physical objects in the indirect realist theory of perception, according to which we only ever directly perceive our own sense-data. Just as on that view physical objects are forever concealed behind a 'veil of perception', here our mental states are concealed behind a veil of introspection.

We may also appreciate a respect in which this causal inferentialist thesis forces a major departure from how we ordinarily think and talk about our own mental states. For considering affective intentional states like emotions for a moment, we ordinarily think of these, not as 'posits', but as things that we *experience*, and by this we do not mean experience just their effects, but *the emotions themselves*. We frequently describe ourselves as *feeling* love, guilt, fear or fury and as *knowing what these are like*, and it would grossly misrepresent these claims to say that what we really mean is that we feel or know only the effects of these things. For this would leave open the question, 'But what is the fear like itself?', a question that those who have 'felt fear' never think of asking.

6.2 Mereological inference

I want to suggest that there is a non-causal sort of inference that might serve as a more appropriate model for understanding inferential self-knowledge. A simple illustration unrelated to self-knowledge would be this. Suppose that a dog is obscured behind a wall except for its tail, and you judge that it belongs to a dog. It seems here that you *infer* that there is a dog there from seeing the tail. However,

[6] Quassim Cassam goes further than Lawlor and even suggests that our knowledge of our own internal promptings–the inner events from which we supposedly infer our own standing mental states–is also inferential. Thus ordinary self-knowledge of our standing attitudes is for Cassam remarkably complex, being the result of inferences from inferences!

this is not inferring from effect to cause or vice-versa. The tail is not an effect or a cause of the dog, but is a part of it. This inference needn't have been explicitly and laboriously articulated in conscious thought: on seeing the tail one might just immediately think, 'There's a dog,' and not, 'There's a dog's tail, so there's a dog.' But the fact that we made an inference would become manifest if our expectation were suddenly disappointed, if we discovered, say, that by some misfortune the tail had become detached from its erstwhile owner.

Note how this case shows that the mediate/immediate and direct/indirect distinctions are logically distinct. Your knowledge that a dog is there was based on knowledge of another fact, namely, that a dog's tail is there, so the knowledge was mediate. But the knowledge was nevertheless *direct* since it was not based on knowledge of *another object*. You inferred the presence of a dog while, and due to, directly perceiving the dog, albeit just a part of it.

But is this really a case of inference? Perhaps some will try to distinguish these kinds of judgments from inferences, saying that instead they are cases of 'seeing as': one does not infer the presence of a dog from the tail but rather sees the tail as the tail of a dog. For some philosophers have maintained that evidence must be distinct from what it is evidence of (Austin 1964: 115; Kelly 2014: §3). A thing, it is insisted, cannot be evidence of itself. True enough, but how about a part of a thing? Can't a part be evidence for the existence of the whole? If so, then evidence isn't necessarily an *intermediary*. Granted, we could imagine variations of the dog case where we would feel less comfortable saying that it involves inference from evidence, such as if half of the dog was in view or the whole dog (we still, in a sense, only see its facing side). But we can also think of cases where speaking of inferring from part to whole seems quite natural, such as when we infer the existence of a nearby ant colony from seeing some colony members out on patrol. Either way, it will at least be admitted that these judgments are inference-like, and, whatever we want to call them, we may explore whether they could serve as a useful model for understanding self-knowledge.

The objection might also be made that inference is always between propositions, but parts and wholes are not propositions. But this complaint would also apply to the causal conception of inference, since causes and effects are not propositions either. The point, however, can be conceded and is nothing to worry about, since inference-involving propositions can be constructed that refer to the parts and wholes, or the causes and effects. We are merely speaking in a loose or contracted manner when we speak of inferring from part to whole or from effect to cause.

Now concerning the internal promptings from the Katherine case mentioned before, Quassim Cassam says, 'These promptings serve as *evidence* that she wants another child, and she knows that she wants another child on the basis of this evidence. There is no plausible sense in which her self-knowledge is "direct"' (2017: 730). However, if we see these promptings not as *effects of* but as *partly constituting* wanting another child, and hence see the inference as being from part to whole, then there *is* a sense in which her self-knowledge is direct: much the same sense as was evident in the dog case. And why not see things that way? Indeed, Lawlor herself gives some credit to such a viewpoint in a footnote: 'Why then do I report my desire as immediately known? It seems likely that this is because … these experienced imaginings are experienced as so intimately bound up with *having* the desire that they aren't experienced as ever having been independent symptoms, by means of which the self-ascription was made' (2009: 68).

However, our case for rejecting the causal conception of these inferences for the mereological one is not just that, as the last quotation suggests, it's more faithful to how things seem. Rather, it is primarily based on the theoretical benefits: by doing so we can accept that some ordinary self-knowledge is evidence-based, thus creating logical space for MDSD, *while preserving the subject's direct relation to the mental state*[7] and so avoiding the alienation objection. When Romeo feels a longing for the absent Juliet, what he is experiencing, what he feels, is *love itself*, albeit in just one of its manifestations. It is not a symptom or effect of love that he feels. Thus the mental state is not something he is at some remove from. It is part of his being. We can have inferential self-knowledge without any alienating veil of introspection.

True, we may say things like, 'He felt a longing for her because he loves her', and some would take this to signal a causal explanation. But this interpretation isn't obligatory, since we often use 'because' where the explanation isn't causal, such as in, 'He made the outburst because he's an irritable

[7] I use this 'direct relation' phrase with reservations, since although common it is highly misleading. This, along with 'direct access' talk, suggests a dualistic picture of there being two entities standing in some relation to each other, a self, and a thought or feeling say, a picture that the grammar of subject-object sentences like 'I feel an itch' suggests to us. This leads us to seek an explanation of the self-knowledge, an account of how this subject and object manage to 'connect'. However, as Hume noted (1978[1739]: bkI. ptIV, §VI), it's entirely obscure what the 'I' is supposed to refer to here. An alternative approach is to reject the dualistic picture. We could do this by conceiving of the mental state simply as a modification of our consciousness. After all, our 'stream' or 'field' of consciousness must be some way or other (just as a physical object must be some shape or other), so why not think of our (conscious) mental states as the way that our consciousness is? One might insist on asking, 'But how does consciousness know what state it is in?' but this can be dismissed as a misbegotten question. It is of the essence of consciousness to 'know itself'. The question attempts once again to impose a dualistic picture, this time between consciousness and its own form, as if they were separate but interacting entities.

person' (such an explanation can be informative, since outbursts are not always manifestations of an irritable character). The former proposition is explanatory in the sense of placing the longing in a pattern, helping us to see it as *part* of a broader phenomenon.

The idea of mereological inference also allows us to explain why self-knowledge is sometimes inferential and sometimes not. For this could be because certain mental states are simple or unstructured while others are structured. With the former, we can be immediately aware of the whole, so no inference is needed (knowing that you have a pain is a likely example), but where the mental state has structure and we are aware of only a part, inference is required.

7. Explaining MDSD

Having shown that inferential self-knowledge can be transparent and direct, an inferentialist view of self-knowledge has been made more attractive. The possibility of MDSD occurring in the normal course of events can now be accepted without this imposing upon us a distorted, 'alienated' picture of our relation to our own mental states. However, though understanding how MDSD is possible in the abstract is all well and good, it would be desirable to have a grasp of some of the more specific ways it can happen. How, for instance, do the mentioned kinds of evidence get 'mistreated'? I will end, then, by attempting to describe, albeit in a sometimes sketchy manner, a number of these processes.

As noted, certain mental states are structurally complex, manifesting in specific ways in different circumstances. This suggests that a way of knowing that we have a particular mental state can include *recalling* how we reacted in such circumstances previously, or *trying to predict* how we would react in those circumstances, by imagining ourselves being in them perhaps. These two methods were suggested by Wittgenstein when he described a process of introspection as involving 'the calling up of memories, of imagined possible situations, and of the feelings that one would have if …'. We could also be in one such circumstance presently, and could note our reaction. Inferences can then be made from the results.

But these routes to self-knowledge are subject to failings, failings of *selective or distorted remembering and imagining*. For instance, when remembering certain things, like a past relationship, we might have many occasions to choose from, and one might remember it as being better (like in nostalgic thinking) or worse than it was, focusing on the good and ignoring the bad or vice versa (see Lawlor 2003: §7), or one's view of a past situation might be coloured by one's current affect

(projecting our current mood into the past situation perhaps). Further, motivation could be behind such distortions: one might focus on the bad in a past relationship in an effort to justify the loss of it (so-called sour grapes), and end up believing that one's feelings for the person are weaker than they in fact are.

Besides collecting evidence selectively, one can attribute some given piece of evidence to the wrong thing. This sort of error is possible because certain phenomena might be associated with different psychological conditions; there could be *qualitative overlap or similarity* between the mental state one is in and others. Is what Romeo feels a part of love, or of a less enduring infatuation, or lust, admiration, friendship, dependency, or obsession?[8] A number of these conditions might involve missing the person or having them stuck in one's mind. Does he even understand the difference between these things? (It is easier to deceive yourself about something when you're left to define the thing yourself.) It is plausible to understand Anna Karenina's self-deception along these lines, where she attributed her amorous feelings not to love but to some kindred condition like friendship.

Misattributing, say, missing S to the wrong condition might occur because one mightn't be clear about *why* exactly one misses S, and what one's reasons for missing S are could determine which condition that belongs to. Does Romeo miss Juliet for her 'intrinsic' qualities (and which ones?), or for more extrinsic factors, like the lifestyle he has with her, or her nice apartment? Or does he just miss having someone, and she is a someone? Most likely his missing her is overdetermined by numerous factors, but can he be sure which has the most weight? Knowing this might require good predictive understanding of counterfactual scenarios (e.g. he thinks that if she didn't have a nice apartment he'd still miss her just as much), which can be difficult.

Some other factors can be mentioned that facilitate MDSD, factors not specifically related to the idea that inferences from evidence are occurring.

One such factor is *vague or indeterminate boundaries* between mental states. Bishop Butler noted that the lack of a determinate boundary between two things allows us to self-servingly imagine it to be where we like:

> There is not a word in our language, which expresses more detestable wickedness than *oppression*; yet the nature of this vice cannot be so exactly stated, nor the bounds of it so

[8] See (Kirsch 2020) for an instructive case illustrating this.

> determinately marked, as that we shall be able to say in all instances, where rigid right and justice ends, and oppression begins. In these cases there is great latitude left, for every one to determine for, and consequently to deceive himself (2006[1726]: 106; also see Sloman *et al*. 2010: 270).

Such vague boundaries exist between mental states, both intentional and non-intentional, and can also facilitate self-deception. Love, for instance, is something that comes in degrees (Brogaard 2015: 170), and it is indeterminate what the boundary between, say, liking and loving someone is. And it is plausible to suppose that people who, for whatever reason, want to see themselves as being or not being in love with someone else might self-servingly imagine that line to be wherever it suits. This also suggests a way in which one can self-deceive about whether one believes something, for how much confidence must one have in it before one will say that one believes it? One might adopt strict or lenient standards as is convenient.

Relatedly, judgments can also be in error from being based on, let's call it, an *inappropriate standard*. Many kinds of judgments are made relative to standards. For instance, Jones, who lives in a small town, thinks that he is hard-working. But when he moves to a big city where life is more fast-paced, he realizes that he isn't so hard-working after all, not compared to these city-folk. Similarly, Romeo might get to know another woman who he becomes completely besotted with. His feelings for Juliet might seem tame in comparison, so much so that he might think that only now does he know what love is really like, and that his love for Juliet wasn't 'true' love. Moreover, it's possible that when one asks oneself the question 'Do I love him/her?' one might self-servingly seize upon a standard for answering it that's more likely to yield the desired answer.

Erroneous beliefs about our own mental states can also arise from having an *impoverished or mistaken understanding of the intentional object* of the mental state. Romeo might believe that he loves Juliet, but he really doesn't, because this Juliet is largely a creation of his own fancy. She is not the woman he thinks she is, and what he loves is a phantasm created by his hopes. Similarly, Flo might think that she believes in some fine-sounding political slogan. But she really doesn't, because if she thought it through she'd realize that she doesn't accept its implications.

None of these ways of being mistaken or self-deceived about one's own mental state requires that it is unconscious and inaccessible to introspection or that the subject is alienated from it. They indicate how MDSD can occur in the normal course of events.[9]

References

Alston, W. (1983) 'What's wrong with immediate knowledge?', *Synthese*, 55/1: 73-95.

Austin, J. L. (1964) *Sense and Sensibilia*. Oxford: Oxford University Press.

Barz, W. (2019) 'The puzzle of transparency and how to solve it', *Canadian Journal of Philosophy*, 49/7: 916-935.

Bem, D. J. (1972) 'Self-Perception Theory', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, vol. 6, 1-62. New York: Academic Press.

Boyle, M. (2019) 'Transparency and reflection', *Canadian Journal of Philosophy*, 49/7: 1012-1039.

Boyle, M. (2015) 'Critical study: Cassam on self-knowledge for humans', *European Journal of Philosophy*, 23/2: 337-348.

Brogaard, B. (2015) *On Romantic Love: Simple Truths about a Complex Emotion*. Oxford: Oxford University Press.

Butler, J. (2006[1726]) 'Upon self-deceit', in D. E. White (ed.). *The Works of Bishop Butler*, 103-109. Rochester: University of Rochester Press.

Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press.

Cassam, Q. (2017) 'What asymmetry? Knowledge of self, knowledge of others, and the inferentialist challenge', *Synthese*, 194/3: 723-741.

Cassam, Q. (2014) *Self-Knowledge for Humans*. Oxford: Oxford University Press.

Coliva, A. (2016) *The Varieties of Self-Knowledge*. London: Palgrave.

Doucet, M. (2011) 'Can we be self-deceived about what we believe? Self-knowledge, self-deception, and rational agency', *European Journal of Philosophy*, 20/S1: e1-e25.

Fernández, J. (2013) *Transparent Minds: A Study of Self-Knowledge*. Oxford: Oxford University Press.

[9] For their helpful comments on earlier drafts of this paper I'd like to thank Mathieu Doucet, Ivan Ivanov, Johannes Roessler and Winnie Sung. The paper also benefited from audience comments during seminars hosted by Sun Yat-sen University, Zuhai, Xiamen University, and the Asian Epistemology Network in 2021.

Hume, D. (1978[1739]) *A Treatise of Human Nature*. P. H. Nidditch, ed. Oxford: Oxford University Press.

Kelly, T. (2014) Evidence. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/evidence/

Kirsch, J. (2020) 'Interpreting our emotions', *Ratio*, 33/1: 68-78.

Lawlor, K. (2009) 'Knowing what one wants', *Philosophy and Phenomenological Research*, 79/1: 47-74.

Lawlor, K. (2008) 'Knowing beliefs, seeking causes', *American Imago*, 65/3: 335-356.

Lawlor, K. (2003) 'Elusive reasons', *Philosophical Psychology*, 16/4: 549-564.

Lynch, K. (2017) 'An agentive non-intentionalist theory of self-deception', *Canadian Journal of Philosophy*, 47(6), 779-798.

Lynch, K. (2016) 'Review of *Self-Knowledge for Humans*, by Quassim Cassam', *Dialectica*, 70/1: 113-119.

Lynch, K. (2012) 'On the "tension" inherent in self-deception', *Philosophical Psychology*, 25(3), 433-450.

Moore, G. E. (1929) 'Indirect knowledge', *Proceedings of the Aristotelian Society, Supplementary Volumes*, 9: 19-50.

Moran, R. (2001) *Authority and Estrangement: An Essay on Self-Knowledge*. Princeton; Oxford: Princeton University Press.

Nussbaum, M. (1988) 'Love's knowledge', in B. P. McLaughlin & A. O. Rorty (eds) *Perspectives on Self-Deception*, 488-514. Berkeley etc.: University of California Press.

Proust, M. (1982[1925]) *Remembrance of Things Past*, vol. 6. Trans. C. K. F. Moncrieff. Vintage.

Ryle, G. (2009[1949]) *The Concept of Mind* (60th anniversary edition). London; New York: Routledge.

Sloman, S. A., Fernbach, P.M., and Hagmayer, Y. (2010) 'Self-deception requires vagueness', *Cognition*, 115/2: 268-281.

Tolstoy, L. (2014[1878]) *Anna Karenina*. Trans. M. Schwartz. New Haven; London: Yale University Press.

Wittgenstein, L. (2009[1953]) Philosophical Investigations, 4th ed. Trans. G. E. M. Anscombe, P. M. S. Hacker & J. Schulte. Sussex: Wiley-Blackwell.