(This is a pre-print of a paper published in *Philosophy*, 2022.)

The pragmatic hypothesis testing theory of self-deception and the belief/acceptance distinction

Kevin Lynch

Abstract

According to the pragmatic hypothesis testing theory, how much evidence we require before we believe something varies depending on the expected costs of falsely believing and disbelieving it. This theory has been used in the self-deception debate to explain our tendencies towards self-deceptive belief formation. This article argues that the application of this theory in the self-deception debate has overlooked the distinction between belief and acceptance, and that the theory in all likelihood models acceptance rather than belief, in which case it is probably not relevant to the explanation of self-deception. It is suggested, however, that doxastic error costs might be relevant to explaining some types of self-deception, though they feature in an evolutionary explanation of it rather than a psychological one.

1. Introduction

A common view among theorists of self-deception is that this condition involves having a belief that is typically false and that goes against the evidence that the subject knows about or could easily have known about. A key question for such theorists is *how* these normally rational subjects manage to have these irrational beliefs. This is sometimes called the question as to 'the dynamics' of self-deception.

The aim of this article is to evaluate a theory on the dynamics of self-deception, one that is strongly associated with the deflationist/non-intentionalist perspective due mainly to the promotional efforts of Alfred Mele (2001) and Dion Scott-Kakures (2002; 2000). This sophisticated view, which has its origins in the psychology literature, attempts to explain belief-formation and doxastic biases as being influenced by prudential desires to avoid certain expected costs, in particular,

the costs of believing falsely. Scott-Kakures (2000), following the psychologists Y. Trope and N. Liberman (1996), calls it the 'pragmatic hypothesis testing model' (PHT model), and Mele calls it the 'FTL theory' or 'FTL model' (an acronym referring to the psychologists J. Friedrich, Y. Trope, and A. Liberman).[1] I will use the former designation since it is more descriptive. Mele has other ideas on the processes that cause self-deceptive beliefs and promotes the PHT model as only a partial explanation, but his frequent use of it to explain particular cases suggests that it is of central importance in his theory of self-deception. Scott-Kakures seems even more enthusiastic about it. The theory has also been employed in other philosophical debates (e.g., Sullivan-Bissett, 2017).

A big selling point of the PHT theory is that it promises a unified explanation of both straight/positive self-deception, where the subject irrationally believes something he wants to be true, and twisted/negative self-deception, where he irrationally believes something he does not want to be true (Mele, 2001, p. 98). This has been an aspiration and significant challenge for theorists of self-deception, and a stumbling block for certain theories. Though other theorists have offered different unifying accounts, the PHT solution certainly seems like an elegant one.

The thesis of this paper is something that I briefly suggested before but did not properly argue for (Lynch, 2017, p. 793): that the PHT model should be examined in light of the concept of acceptance, and that in all likelihood what it offers is a model of acceptance formation rather than belief formation. But if self-deception concerns belief (as PHT theorists themselves assume), then this would imply that the PHT model is not explanatorily relevant to self-deception. Towards showing this, in the next two sections the PHT theory and then the belief/acceptance distinction are explained. I then show that cases remarkably similar to those discussed by PHT theorists turn up in the literature on acceptance. The paper's longest section then follows, which argues that it is more plausible to understand PHT cases in terms of acceptance rather than belief. Afterwards, a revisionist response is discussed according to which self-deception results in acceptance, and then the paper ends with an alternative suggestion about the explanatory relevance of doxastic error costs for self-deception. As we will see and as has not been appreciated by theorists of self-

[1] Scott-Kakures (2000) traces the idea to some earlier papers by M. Lewicka published in the *Polish Psychological Bulletin*.

deception, the PHT theory raises some fundamental questions about the nature of belief, and it is of no small relevance to other debates such as between evidentialists and pragmatists, and doxastic voluntarists and involuntarists.

2. The pragmatic hypothesis tester theory

It is not controversial to say that people normally need evidence or reasons before they will believe something. But *how much* evidence is needed? The PHT theory asserts that the amount can *vary*, and it represents this variation in the key idea of a *confidence threshold*, which refers to the quantity and quality (I'll just speak of the 'amount') of evidence required before one comes to believe some proposition, *p*. Furthermore, according to the theory, these thresholds get set at a certain level mainly depending on the magnitude of expected *costs*.[2]

The first of these costs is the *cost of acquiring the evidence*, that is, how much time, money, or energy must be expended to acquire it (Mele, 2001, p. 35). Where evidence concerning whether *p* is very costly to acquire, subjects might require less of it before judging (which instigates believing) that *p*/not-*p* compared to situations where such evidence is less costly to acquire. Furthermore, the *significance of the issue* for the subject can be an interacting factor here; if the issue is not so important, a subject might be content to make a judgment after only a cursory investigation, while for a more important or interesting matter she might inquire more deeply (Scott-Kakures, 2000, p. 363).

The costs which do the most important explanatory work, however, are the *costs of falsely believing that p/not-p*. Believers are here theorized as pragmatic, cautious thinkers concerned more with avoiding costly mistakes than simply with knowing truths. If the expected costs of falsely believing that *p* are low, then lax standards can be afforded and one's confidence threshold for believing that *p* will be low (that is, it will take less evidence before one believes it). On the other hand, if the expected costs of falsely believing that *p* are high, then one will be more circumspect and one's confidence threshold will be high (it will take a greater amount of evidence to make one believe it). Importantly, one's confidence thresholds can be *asymmetric*

[2] Presumably it would also depend on the perceived likelihood of the costs obtaining, though PHT theorists rarely mention this. We can ignore this point for current purposes and assume that all costs are regarded as equally probable.

in magnitude when judging whether *p* or not-*p*. For instance, due to different cost expectations one might have a lower threshold for concluding that *p* than for concluding that not-*p*, in which case one will be *biased* towards believing that *p*. Cost differences will also lead to differences in how subjects treat evidence. High costs associated with falsely believing that *p* would lead to greater scrutiny of evidence for *p*, since one wishes not to be misled by it.

Mele's case of Gordon usefully illustrates how this theory might be used to explain both straight and twisted self-deception. Gordon works for the CIA and is under suspicion for being a double-agent. Both his parents and his colleagues in the agency don't want this to be true, and suppose that they have roughly the same evidence about the matter. But the potential costs of believing this if it's false are very different for the two parties. For Gordon's parents, there would be no obvious cost with falsely believing that Gordon is innocent, though there would be costs associated with falsely believing that he is guilty (Gordon might feel abandoned and betrayed so their relationship with him could be damaged, and they would also suffer mental distress in believing this). So Gordon's parents will be biased by these cost aversions towards believing the *welcome* proposition that Gordon is innocent. Things are different for Gordon's colleagues however. For them, falsely believing that Gordon is innocent and carrying on as per usual could expose them to significant risks. So they will be biased towards believing the *unwelcome* proposition that Gordon is guilty.

Note that this case is supposed to illustrate, not straight and twisted self-deception, but just the motivational complexes that might be (partly) responsible for them. Mele does not claim that either Gordon's parents or his colleagues are self-deceived, and they could hardly both be given that they believe opposite things. The parents and colleagues supposedly have beliefs about Gordon that are motivationally biased, but Mele would require further conditions to be satisfied before they would be self-deceived. Specifically, their beliefs would also need to be false, and perhaps also, the evidence in their possession would need to provide greater warrant for the contrary conclusion (2001, p. 120).

Furthermore, aversions towards the costs of falsely believing things might not be enough even just as far as motivational states go. In straight self-deception, the subject is supposed to not only believe something that he wants to be true, but also *because* he wants it to be true: the desire that *p* causes the belief that *p*. Thus Mele is concerned to show an *interaction* between the desire that *p* (e.g., that Gordon is

innocent) and the desire to avoid the costs of falsely believing that *p*/not-*p*. The idea here is that the fact that state-of-affairs *p* is desired is cost-generating. This is because if one desires that *p*, an immediate cost of falsely believing that not-*p* will be 'considerable emotional distress' (Mele, 2001, p. 36. Also see Friedrich, 1993, p. 314). Believing things that we really don't want to be true is typically upsetting. Curiously, the causal status of the desire that is definitionally essential for twisted self-deception is less clear (nobody says that twisted self-deceivers believe that *p* *because* they want that to be *not* true), which makes it more puzzling.

Mele's emphasis on the cost of negative affect, however, undermines his idea that the PHT theory provides a unified explanation of straight and twisted self-deception. For note that Gordon's parents suffering this emotional distress is not dependent on the unwelcome belief being false. The distress arises simply from their believing it. (Similarly, a sinful Christian's belief that God will punish him might cause him anxiety, whether such a God exists or not. The belief-state can cause the anxiety independently of its truth-value.) Therefore, this is not a cost of *being in error or mistaken* (see Sarzano, 2018, p. 109). It is therefore doubtful that we can consider this explanation to be an application of the PHT framework, which primarily explains in terms of *error*-costs (e.g., Friedrich (1993) calls his theory the 'PEDMIN analysis', which stands for Primary Error Detection and Minimization, and Trope *et al*'s key explanatory concepts are the 'costs of false acceptance and false rejection of a hypothesis' (1997, p. 112)). Perhaps the desire that *p* generates genuine error costs in other ways though, and I will ignore this problem from here on so as to focus on a more fundamental difficulty with the PHT theory.

3. The distinction between belief and acceptance

Both Mele and Scott-Kakures assume that the thresholds mentioned above are belief thresholds. However, they habitually describe them as 'acceptance' and 'rejection' thresholds. But they give every indication that they are treating 'accept' as a synonym for 'believe' and 'reject' as a synonym for 'disbelieve'. This, however, is not necessarily in line with the original intent of the PHT theory. The psychologist James Friedrich, one of its originators, did not explicitly present it as a model for understanding belief formation (or acceptance formation for that matter) but rather as a 'model of lay hypothesis testing' (1993, p. 300), which might suggest that the

thresholds just represent the points at which hypothesis testing is discontinued. But in the work of Trope *et al* (1997) they are clearly represented as belief thresholds, though they are also frequently called 'acceptance/rejection thresholds'.

Some philosophers, however – concerned as they often are with conceptual or classificatory nuances – have distinguished acceptance/rejection from belief/disbelief. Though 'accept' in ordinary English is indeed sometimes used interchangeably with 'believe', it also has other uses. Accepting that *p*, as understood by these philosophers, is supposed to include things like taking it for granted that *p* (Bratman, 1992, p. 4), positing or postulating that *p*, where one then acts on that assumption or makes inferences on its basis (Cohen, 1992, p. 4), working on the assumption that *p* (Davis, 1988, p. 175) or presuming it to be true that *p* (Ullman-Margalit, 1983). Thus it is not necessarily a unified attitude (Engel, 1998, p. 140).

To show that some type X is distinct from some type Y we should show cases of X existing without Y, and Y existing without X. So let's look at putative cases where one accepts that *p* but does not believe that *p*, and cases where one believes that *p* but does not accept that *p*.

First, here are some cases of acceptance without belief. One would be the case of a lawyer who, due to professional obligations, accepts that his client is innocent though he believes, or strongly suspects, that he is guilty (Cohen, 1992, p. 20). Davis (1988, p. 174) mentions the case of Galileo, who publically accepted geocentrism and rejected heliocentrism at the behest of the Inquisition, though he privately maintained belief in heliocentrism. Another is the case of a scientist who accepts a theory by committing herself to its research programme, despite serious anomalies and while believing that, like most scientific theories, it will probably be replaced by a better one in the future (Cohen, 1992, p. 90; van Fraassen, 1980, p. 12). Blaise Pascal urged his readers to accept Catholic doctrine despite their agnosticism or atheism for pragmatic reasons (Cohen, 1992, p. 18), and a teacher might not believe that a student's essay is his own work, but she might accept that it is for want of proof of plagiarism.

Cases of belief without acceptance seem less common in the literature, but one might be where a juror acquires a belief about a defendant outside of court, but dismisses it for the purposes of reasoning about the defendant's guilt in court, despite the belief being relevant to that question (Cohen, 1989, p. 369–370). And as a further illustration, a judge might believe that a defendant is guilty of a crime but might still

not accept him as guilty (finding him not guilty in her official judgment) because the prescribed standards of proof have not been reached (Cohen, 1992, pp. 122–124).

These cases illustrate the general distinguishing features that are attributed to acceptance (see Bratman, 1992, pp. 3–4; Engel, 1998, pp. 146–147).

1) Acceptance, unlike belief, is *voluntary or intentional*. This is evident with the lawyer defending the guilty client; if his scruples got the better of him he might have declined to defend him and accept him as innocent. He chose to defend that client, with what that entails. In contrast, the evidence of his client's guilt forced the belief upon him. Also, accepting must be voluntary if the Inquisitor expected Galileo to do it on his insistence.
2) Acceptance is also said to be *context-dependent*. We see this with Galileo, who accepted geocentrism in public but not in private, though he believed the same in public and private. We also see it with the juror who rejected a relevant belief acquired outside of court when reasoning in court.
3) It is said that, unlike belief, acceptance *does not 'aim at truth'*, but at utility or some other value, and *needn't be shaped by evidence*. We see this with Galileo again, who accepted geocentrism not because he thought it was true but *with the aim of avoiding punishment* from the Inquisitor. It was also accepted not on the basis of evidence but *on the basis of threats*. The threats were his *reason* for accepting it.
4) Acceptance is *not regulated by an ideal of rational integration with other attitudes* in the same sense as belief. For instance, the judge who rejected the defendant's guilt (of financial fraud, say) for want of particular standards of proof being met might still not be trusting towards that person (she wouldn't be inclined to invest in his latest investment scheme), though someone who genuinely believed the defendant was innocent would probably have more congruent trusting attitudes.
5) Acceptance is said to be *an all-or-nothing affair* in contrast with belief, which can come in different strengths. We see this in the legal cases above where acceptance or rejection is expressed in findings of guilty or not guilty, which admit of no degrees or qualifications. Beliefs, on the other hand, can be held with various degrees of confidence.

There may well be important differences between the cases mentioned above, which might suggest that 'acceptance' denotes a family of attitudes. For instance, in some cases the accepting is a kind of assuming or taking for granted while in others it is not, and in some it is closely tied to a speech act and outward feigning of belief but in others less so. I don't think we should worry too much about this, since the category of belief too is variegated (think of how tacit beliefs, delusional beliefs, some religious beliefs and perceptual beliefs differ from paradigm cases of belief). At any rate, features (1) to (5) mostly do seem to apply to the examples of acceptance looked at, giving them some degree of unity. Let us proceed to see how this distinction bears on the PHT phenomenon.

4. The PHT model and the acceptance literature

A key point I want to make here is that the explanatory apparatus of the PHT theory appears in the philosophical literature on acceptance and has been used to explain how people come to accept things or not, in particular in Bratman's seminal 1992 paper. In a section of that paper entitled 'Asymmetries in the costs of errors', Bratman describes a number of cases that look strikingly similar to those used to illustrate the PHT theory, the first of which is this:

> I am planning for a major construction project to begin next month. I need to decide now whether to do the entire project at once or instead to break the project into two parts, to be executed separately. The rationale for the second strategy is that I am unsure whether I presently have the financial resources to do the whole thing at once. I know that in the case of each sub-contractor – carpenter, plumber, and so on – it is only possible at present to get an estimate of the range of potential costs. In the face of this uncertainty I proceed in a cautious way: In the case of each sub-contractor I take it for granted that the total costs will be at the top of the estimated range. On the basis of these assumptions I determine whether I have at present enough money to do the whole project at once (1992, p. 6).

It seems that expected error costs are being factored into this person's reasoning. Two mistakes are possible in his estimation of the project's financial cost: underestimation

or overestimation. If he underestimates it, he could be left with an unfinished project that he has sunk all his money into but that is generating no return, which could lead to bankruptcy. There are no costs of similar magnitude with overestimating the financial costs. Since he wishes to avoid the more serious error of underestimation, he 'take[s] it for granted' that the costs will be at the top of the estimated range. But this is not to say he believes it will cost that much, and Bratman continues, 'In contrast, if you offered me a bet on the actual total cost of the project—the winner being the person whose guess is closest to the actual total—I would reason differently'.

Perhaps the property developer is reasoning in a more explicit fashion than what PHT theorists had in mind, who do 'not presume that such computations generally take place in a kind of conscious, thoughtful manner' (Friedrich, 1993, p. 317).[3] But another case used by Bratman is even closer to one used to illustrate the PHT theory. This case is of when you are driving on a narrow winding road, and you 'assume' (p. 6, note 11) that there will be a car coming towards you on the blind turns, even if you believe this is unlikely. This is very similar to Scott-Kakures' case of Olga, who is biased towards 'accepting' (2000, p. 364) that other drivers she meets are bad drivers after the dangers of driving were made salient to her by some recent near-accidents. (My driving instructor once told me to always assume that the other drivers are idiots).

Despite these notable similarities, Mele and Scott-Kakures make no reference to the work of Bratman, and pay no heed to the distinction between belief and acceptance. Though the words 'accept' and 'reject' roll naturally off their tongues when they describe the PHT theory and their illustrative cases, they treat these as synonyms for 'believe' and 'disbelieve'. This gives rise to a pressing question: Why should we think that the PHT model is a model of belief rather than of acceptance?

5. Being shaped by non-evidential reasons

Let's consider Mele's Gordon case again, focusing on the colleagues rather than the parents (since as was said, it's not clear that the parents were worried about *error*-costs), and let's see if it can be plausibly understood in terms of acceptance. Let the

[3] These authors, however, give no reasons for thinking that these calculations generally *could not* take place in a conscious, thoughtful manner.

word 'assume[tt]' be a technical term that's ambiguous between meaning believing or accepting. We can then say in neutral fashion that Gordon's colleagues have a bias towards assuming[tt] that Gordon is guilty as charged. Our question is whether their assuming[tt] is a case of believing or of accepting.

It certainly seems possible that this was a case of accepting. The colleagues might have believed that Gordon is probably innocent, though they might nevertheless have decided to 'err on the side of caution' or 'play it safe' and tentatively treat him as guilty. There is nothing that rules out that interpretation at least. But as I will argue, if we take seriously the distinction between belief and acceptance then we have positive reasons to favour that interpretation. This is based on the third characteristic of acceptance mentioned above, which includes two intimately related points: acceptance, unlike belief, 'aims not at truth, but at utility or success' and 'need not be shaped by evidence or evidential reasons' (Engel, 1998, p. 146).

Firstly, the attitude of the colleagues does not seem to 'aim at truth' specifically. Instead it aims at *the avoidance of certain harms or evils*, a prudential aim.[4] Remember that the PHT theory's claim is not simply that the subject wants or aims to avoid being in error about some issue, which might just be part-and-parcel of aiming to ascertain the truth about it. Rather, the claim is that the subject aims to avoid *the costs* of being in error. This is the key explanans of the theory and the posited motive behind the assuming[tt], and it is a different aim from the aim to avoid error *simpliciter*. (To appreciate the difference, we should note that for us humans, error can be an ultimate evil just as knowledge can be an ultimate good. Pure curiosity is a basic human motive irreducible to others. The costs associated with being in error are an additional and extraneous evil to just being in error.) So the 'aim' of the assumption[tt] is one we would more associate with accepting rather than believing, given the standard way of distinguishing those.

Second, it is said that acceptance needn't be shaped by evidential reasons, and this is supposed to distinguish acceptance from belief, the implication being that beliefs can *only* be shaped by evidential reasons. Now it certainly is true that

[4] Many consider talk of belief having an aim to be metaphorical (see Fassio, 2015; Shah and Velleman, 2005, pp. 498–499; Wedgwood, 2002, p. 267). But given that accepting is supposed to be voluntary, and something that we do, it could be taken to have an aim in a more literal sense (or better, *we* could be taken to have an aim in doing it). Talk of the 'aim' (or perhaps aims) of acceptance should be less problematic than talk of the 'aim of belief'.

according to the PHT theory the colleagues' assumption[tt] is shaped, in *some* sense of 'shaped', by something other than evidential reasons, in particular, by expected error-costs. So does that mean it must be an acceptance, granting this way of distinguishing acceptance from belief? Not necessarily. The term 'shaped', which was used by Bratman, is vague (deliberately so perhaps) and could mean different things. And on some interpretations, many would agree that beliefs can indeed be shaped by non-evidential factors like error-cost expectations, for instance, if 'shaped' simply meant 'caused'. So we must investigate exactly what kind of role these error-cost expectations could be playing in 'shaping' the assumption[tt]. Some such roles might be more congenial to understanding the assumption[tt] as an acceptance, while others might encourage thinking of it as a belief. To that end, I believe that we can distinguish three options for the kind of explanatory or 'shaping' role error-cost expectations could be playing, or rather, two options, with the final one breaking down into two. But I will argue that none of them encourage us to think of the assumption[tt] as a belief.

Option 1) The first possibility we may consider gives most credibility to the claim that beliefs cannot be shaped by error-cost expectations. It is that they 'shape' the assumption[tt] in the sense of being the colleagues' *reasons* for assuming[tt] as they do.[5] They 'shape' the assumption[tt] by entering into the colleagues' deliberations or reasoning about the Gordon matter and making the assumption[tt] that he is guilty appear as reasonable or justified. Thus, if one were to ask the colleagues, 'Why are you assuming[tt] that Gordon is guilty?' or 'Why are you treating him that way?' they might *justify themselves* by mentioning those risks.

Note that I said they'd mention the risks, and not the fact that they expect those risks. By saying that 'the error-costs expectations' might be their reasons for the assumption[tt], I am not suggesting that the *mental state of expecting* is the reason. The word 'expectation' is ambiguous, and can refer to either the mental state or the 'content' of the state, i.e., *what* is expected, which we might characterise as a *future possibility* or *risk*. It's the latter that would be the reason, as it would be the risk itself, and not the subjective fact that they expect the risk, that the colleagues would be

[5] Schroeder (2012) seems to suggest that error-costs can be reasons for belief.

occupied with in their deliberations and that would actually justify their treatment of Gordon.

However, the idea that error-cost expectations are playing a rationalizing role in relation to the assumption[tt] seems more appropriate to the acceptance interpretation of it. For it is implausible that such considerations could enter into *doxastic* deliberation and be taken as reasons for *belief*.[6] Try to imagine an investigator asked to find out whether *p* reasoning as follows: 'This is some evidence for *p* though it's fairly weak. But I don't have the time to look into this much further and there's little at stake if I get things wrong. So I'll just go ahead and believe that *p*.' But if this isn't strange enough, consider that if we agree that the question whether to believe that *p* is transparent to the question whether it's true that *p*, then the reasoning would be, 'This is some evidence for *p* though it's fairly weak. But I'm pressed for time and there's little at stake if I get things wrong. So it's true that *p*'. A converse example of such reasoning would be this: 'The evidence for *p* is very strong. It looks like *p* is true, but it would be disastrous for me if I mistakenly took it for granted, so it's not true that *p*'. This reasoning is fanciful, nay unintelligible (imagine mentioning those reasons to persuade *someone else* that *p*/not-*p*). Note also that we don't need to posit such forms of reasoning to explain how subjects can avoid such error-costs, since the mentioned disaster typically would only follow from *acting* on the assumption that *p*, so it could be avoided by apportioning one's doxastic state to the evidence while simply refraining from acting so (see Jackson, 2019).

But here's a possible rejoinder. Philosophers who say that only truth-relevant considerations can be reasons for belief are often quick to emphasise that they mean that only such considerations can be *consciously* regarded as reasons for belief. But it is sometimes said that people can act, believe, want etc. for reasons of which they are unaware or won't acknowledge. If that's true it raises the question, could error-cost

[6] It is important not to conflate *reasons to believe something* with *reasons to make yourself believe something* here (e.g. see Berker, 2018. Skorupski, 2009). We might indeed acknowledge non-evidential considerations as reasons to make ourselves believe something, i.e. to take action that would, if at all possible, result in us believing it. But as understood here, a reason to believe something is a reason that, once considered, can immediately issue in a belief. That there is a pitter-patter sound on the window is a reason to believe it's raining in this sense, since on hearing it I immediately believe that it's raining. I don't have to do anything, after recognizing that sound, to believe this. The claim is that it is doubtful that error-cost considerations can be reasons to believe in this sense.

considerations be reasons for belief if they are 'unconscious reasons'? The question is, no doubt, a thorny one, and I won't attempt to grapple with it. Indeed, we can grant it a positive answer. For the important point is that nothing in the PHT theory suggests that subjects' expectations or reasoning about error-costs would always be or even tend to be unconscious. Without any positive reasons for holding these to be unconscious, this attempt to defend option (1) is *ad hoc*.

The idea that error-cost considerations can enter into doxastic deliberation and be 'motivating' reasons for belief is controversial at best. What is not controversial, however, is that they could enter into deliberation about whether to *accept* some proposition, since, as we saw in section 4, they are stock examples of the sorts of considerations that rationalize acceptance. Therefore, if the error-cost expectations are playing a rationalizing role in relation to the colleagues' assumption^{tt}, then it is most natural to interpret it as an acceptance.

Option 2) Alternatively, we could suppose that the error-cost considerations 'shape' the assumption^{tt} without being reasons for it, and perhaps this idea is more suitable to interpreting the assumption^{tt} as a belief in these PHT cases. We see an idea like this in how Ema Sullivan-Bissett (2017, p. 728) understands the PHT theory. Sullivan-Bissett doesn't question the idea that subjects' attitudes in PHT cases are beliefs, but she doesn't regard the error-cost considerations as reasons for those beliefs. In fact, she thinks we can reconcile the idea that error-cost considerations influence belief formation with the 'exclusivity thesis': the view that in doxastic deliberation we only accept evidence, or truth-relevant considerations, as reasons for belief. Her idea is that when reasoning about whether to believe that *p* (that is, about whether it's true that *p*), we only consider evidential considerations, and only acknowledge them as reasons to believe. But non-evidential considerations can still influence belief formation, in particular by influencing our confidence thresholds; they can determine *how much* evidence for *p* we will require before we believe that *p*. In her words, 'we can understand these [non-evidential] considerations as functioning to modify the standards for sufficient evidence required for belief, and not as reasons for the subject to withhold or form belief' (2017, pp. 726–727). That is, '[n]on-evidential considerations are not recognized by the subject *as reasons* … but rather, affect the evidential standards required for belief' (p. 727). Or again, 'these considerations merely change the standards required for believing in a particular context, they do *not*

provide non-evidential reasons for forming or withholding belief' (p. 721; Also see Worsnip, 2021, pp. 533–534).

However, Sullivan-Bissett does not tell us exactly how to understand the manner in which the non-evidential considerations 'function to modify' or 'affect the evidential standards'. Two possibilities come to mind.

2a) First, perhaps the idea is that they function as *our reasons for adopting such standards*. On this proposal, the error-cost expectations are indeed functioning as the subjects' reasons, but not reasons *for the belief*; instead, they are reasons *for the adoption of an evidential standard for forming the belief*. So returning to our case, we are to imagine that Gordon's colleagues adopt a standard for belief in a reasoned fashion based on prudential considerations, for instance they think, 'It could be very costly for me if I believe Gordon is innocent when he is not, so I'm going to require strong evidence before I'll believe he is innocent'.[7] Then they proceed to gather and assess the relevant evidence. A belief is then formed on the basis of that evidence.

The following analogy might help us to understand this proposal. Suppose you are designing a lecture course. You deliberate over how high to set the pass threshold. This is done for various pragmatic reasons, like having the right level of difficulty, having an acceptable failure rate, or following an industry standard. Then later, you grade the students' essays, giving a percentage score to each one, and you do that based purely on academic criteria. Your passing/failing the students is a function of both decisions. But the key idea here is that your reasons for passing/failing the students are the academic considerations and not the pragmatic ones, which are only your reasons for setting the pass threshold. Similarly, the colleagues' reasons for their belief about Gordon are simply the evidential considerations about him, not the error-cost considerations, which are their reasons for adopting the evidential standard. Or at best, the error-cost considerations are 'reasons for the belief' only in an indirect sense, only via their being reasons for adopting the standard.

Although the non-evidential considerations are playing a rationalizing role here, the proposal still seems different from option (1). In the fanciful reasoning in

[7] Speaking of confidence thresholds implies a precision, the picture of crossing a precise line, that is purely imaginary and it would be more likely that subjects would think in such vague terms as this. The idea of a confidence threshold is an idealization, like the idea of a point-mass in physics.

option (1), evidential and non-evidential considerations were mixed up in a single episode of deliberation that resulted in a single conclusion. But here the two kinds of considerations are segregated into independent episodes of deliberation, with independent conclusions. This makes the present proposal appear less implausible. But is there really a significant difference between (2a) and (1)?

Consider the analogy again. Here it is quite possible to do the two deliberations in the reverse order. First you could appraise the essays, giving each one a percentage score. Then afterwards you could decide on the pass threshold. Only when you make that latter decision do the students' scores become passes or fails. Likewise, on this proposal it should be possible for Gordon's colleagues to gather evidence about Gordon's guilt/innocence first, while suspending judgment on that evidence since no evidential standard has yet been adopted. Afterwards they decide on what evidential standard to adopt based in part on error-cost expectations, and a belief about Gordon's innocence results. In this circumstance, would the error-cost expectations be reasons for the belief?

I think it becomes more plausible to so regard them when we imagine the decisions occurring in this reversed order. For if the non-academic considerations have a right to be placed among your reasons for passing/failing the students in the reversed analogy case, as seems plausible,[8] then the error-cost considerations have as much of a right to be placed among the colleagues' reasons for believing that Gordon is guilty in the reversed Gordon case. Furthermore, there is one important condition that these error-cost considerations would meet for counting as their reasons for this belief. Earlier (footnote 6) it was suggested that a reason to believe something is a reason that, once considered, can immediately issue in a belief, and that this distinguishes a reason to believe something from a reason to make yourself believe something. But the non-evidential considerations seem to satisfy this condition in the reversed case. The colleagues have gathered their evidence and just need to decide on an evidential standard to convert that awareness of the evidence into a belief. Once

[8] Imagine having a conversation with a student who failed. He asks, 'I thought I did well. Why did you fail me?', and you answer, 'Because this is an advanced course and I've set a high standard for it. I'm afraid your essay didn't reach that standard'. Here you would be mentioning a 'non-academic' consideration in giving your reasons for failing the student. (Of course, you could do this in the unreversed case too.)

that is done the belief immediately forms, just as once you decide on the pass threshold, the essay scores immediately become passes or fails.

It looked like option (2a) was substantially different from option (1), but the differences now appear superficial. It is just a difference of temporal ordering. Thus (2a) inherits all the difficulties of (1), and it cannot rescue the thesis that the assumption[tt] is a belief.

2b) There could be another way to understand how the error-cost expectations affect the evidential standards, however, which is by saying that they are not playing any kind of rationalizing role at all, but operate in a merely causal or mechanistic manner. Sullivan-Bissett claims (in conversation) to have had this in mind. Could this allow us to understand the colleagues' assumptions[tt] as beliefs?

Let us consider more precisely what exactly are supposed to be the causes here. The PHT theory speaks of *error-costs* in its explanations, but they are things that would (potentially) obtain in the future, and therefore they cannot be present causes. But it is *expected* or *anticipated* error-costs (Mele, 2001, p. 35; Scott-Kakures, 2000, p. 363) – which can diverge from actual costs – that are explanatorily important in the theory. So perhaps the causes are error-cost *expectations*, conceived of as propositional things: the expectation *that* such-and-such will (or might) be suffered if I take for granted that *p* when *p* is false. This fits with philosophers frequently describing the relevant determinants as non-evidential *considerations*, since considerations are propositional. However, a proposition or consideration is an abstract thing and also cannot, as such, function as a present cause in a psychological explanation. Considerations can only make a difference in the world by *being considered*, by being intellectually grasped by a subject. And nothing sub-personal, nothing short of an intellect, can grasp and engage with such things as *considerations about future conditional possibilities*, which are the posited determinants of the PHT theory.

But couldn't the cognizing of such considerations affect a subject's confidence thresholds in some merely causal manner? Let me say that I have no argument to rule out this proposal. Certainly acts of thinking do sometimes have purely causal effects, like in associative thinking. But I take it, instead, to be a question of which interpretation of their explanatory role is most natural or plausible, a rationalizing or a merely causal one.

It might be helpful at this stage to look at how these sorts of cases have been understood by philosophers who have had no theoretical investment in our question. To this end, we can make use of an important observation made recently by Jie Gao (2021) and Melanie Sarzano (2018) that PHT cases in the self-deception literature are structurally very similar to 'pragmatic encroachment' cases in the epistemology literature. In pragmatic encroachment cases, error-costs expectations are similarly theorized as influencing knowledge, epistemic justification or belief, as well as action (so yes, we have the same phenomenon independently appearing in the literature on self-deception, acceptance, and pragmatic encroachment!).[9] The focus here is on their epistemic consequences but the interesting question for us is, how has their explanatory role in pragmatic encroachment cases been viewed? Have the pragmatic (error-cost) considerations been viewed as playing a rationalizing role or a merely causal one? Consider the high-stakes version of the original pragmatic encroachment case – Keith DeRose's famous bank case – and how he describes it:

> My wife and I drive past the bank on Friday afternoon … and notice the long lines. I … suggest that we deposit our paychecks on Saturday morning, explaining that I was at the bank on Saturday morning only two weeks ago and discovered that it was open until noon. But … we have just written a very large and very important check. If our paychecks are not deposited into our checking account before Monday morning, the important check we wrote will bounce, leaving us in a *very* bad situation. And, of course, the bank is not open on Sunday. My wife reminds me of these facts. She then says, "Banks do change their hours. Do you know the bank will be open tomorrow?" Remaining as confident as I was before that the bank will be open then, still, I reply, "Well, no. I'd better go in and make sure." (1992, p. 913).

The couple's quandary here looks similar to that of Gordon's colleagues. The couple are reluctant to assume[tt] the welcome proposition that the bank will be open on

[9] Gao and Sarzano say that this observation raises a puzzle, because while pragmatic encroachment cases are thought to exemplify rationality, the PHT cases in the self-deception literature are thought to exemplify irrationality (because self-deception is a kind of irrationality). They each provide interesting solutions to this puzzle. On the present view the puzzle does not arise because we deny that PHT cases are cases of self-deception.

Saturday, just as the colleagues were reluctant to assume[tt] the welcome proposition that Gordon is innocent of the charges against him. And they are both reluctant to assume[tt] these things because of what the expected costs of falsely assuming[tt] so would be. But as the sentence 'My wife reminds me of these facts.' makes clear, these error-cost expectations are playing a rationalizing role in a personal-level explanation. The error-costs are *explicitly discussed* and *contemplated*, and an appropriate response is made. That response is *based on* and *justified by* those considerations, and not just caused by them. It is telling that DeRose – who had no axe to grind regarding our present question – sees things that way, and it would be surprising if error-cost expectations played a fundamentally different kind of explanatory role in the similar-looking Gordon case. Here is another short pragmatic encroachment case:

> Suppose you and your spouse pack up the car and leave for a vacation. On your way out of the driveway, you have the following conversation:
> Spouse: Did you remember to turn the stove off after breakfast?
> You: Yes.
> Spouse: You know you forgot to turn it off the other day. If we leave it on over our vacation, our house will burn down.
> You: You're right. I'd better go back and check (McBrayer 2014).

Here again we have a non-evidential, error-cost consideration ('If we leave it on over our vacation, our house will burn down.'), in conjunction with an evidential consideration ('You know you forgot to turn it off the other day.') causing a change in what is assumed[tt]. The similarity between this and PHT cases is clear. But what is also clear is that the considerations are involved in a rationalizing explanation. Like in the bank case, they are explicitly brought up in an episode of conscious, joint deliberation and the subsequent action (going back to check) is done on their basis.

Sarzano claims that the 'same underlying belief-formation mechanism' (2018, p. 96) is operating in both PHT and pragmatic encroachment cases. We are not taking for granted that it is beliefs that are being formed and would call it an assumption[tt]-forming mechanism for now, but if she is right about the mechanism being the same, and if in pragmatic encroachment cases the error-cost expectations are clearly playing a rationalizing explanatory role, then we have good reason to think that they are also playing a rationalizing explanatory role in PHT cases. And as was determined above,

this idea is more appropriate to understanding the rationalized attitude as an acceptance. I would add here that with regard to the alternative view – that the error-cost expectations are playing a merely causal explanatory role – no account of how this could happen has been developed. This at least shifts the burden of proof onto those who suggest that their explanatory role is a merely causal one.

In summary, we have encountered problems trying to interpret the colleagues' attitudes as beliefs in the Gordon case. First, we tried to think of the error-cost considerations as things the colleagues regard as reasons for their beliefs, but this led to absurdity. Then we considered the possibility that they are reasons for the adoption of evidential standards rather than for the beliefs, but this seemed only superficially different from the first proposal. Then we tried a merely causal interpretation on for size, but this proved unlikely; *considerations* are the kinds of things that have influence by being considered, by entering into our thinking and reasoning and playing a rationalizing role.

Thus if it's best to think of the explanatory role of these error-cost considerations as a rationalizing one, and as rationalizing in relation to the assumptions[tt], and if that fits best with interpreting those assumptions[tt] as acceptances, then we have positive reason for thinking that the attitudes of the colleagues, and of subjects in PHT cases more generally, are acceptances rather than beliefs.

6. The revisionist option

Suppose now that the PHT model is best thought of as a model for acceptance rather than belief. What implications would that have for it as a theory of self-deception? Prima facie, the implications would seem negative. It is a common assumption among theorists of self-deception that self-deceptive processes result in unjustified, irrational *beliefs*, and this is certainly what Mele and Scott-Kakures assume. So insofar as the PHT model is a model for acceptance formation it cannot help us understand self-deceptive *belief* formation/maintenance at all.

But here a dialectical option emerges that might save the relevance of the PHT theory for the explanation of self-deception. We could agree that the PHT theory models acceptance formation and revise our conception of self-deception to

accommodate this: we now hold that self-deceptive processes, or these ones at least, can result in acceptance rather than belief.

I call this the 'revisionist' option, because *PHT theorists themselves* have held that self-deceptive processes result in irrational belief, so they would be abandoning something they previously regarded as a platitude about self-deception. Moreover, this would require a much more significant rewrite of their theories of self-deception than just abandoning their PHT theory, since it is giving up an assumption that is more fundamental and more intimately connected to other assumptions. Thus I wouldn't anticipate them having much enthusiasm for this option.

There is, however, a small number of philosophers who have held that self-deception results in acceptance rather than belief, the first of whom was L. J. Cohen (1992, chap. 5; Also see Frankish, 2004, chap. 8). Cohen did not suggest this in an attempt to defend anything like the PHT theory, but in an attempt to resolve the doxastic paradox of self-deception. For Cohen, straight self-deceivers don't believe that not-*p* while also believing that *p*, but rather they believe (truly) that not-*p* while accepting that *p*.

But the view that self-deception results in acceptance is not only revisionary with respect to the doxastic assumption, but also potentially with respect to the connected *irrationality assumption* about self-deception. Self-deception has generally been regarded as a kind of irrationality or as an irrational state.[10] Furthermore, on doxastic views, there is little difficulty in stating what this irrationality consists in, for instance, we can say it consists in the self-deceiver's belief not being justified by his evidence, or in its resulting from a biased treatment of the evidence. However, if we maintain that self-deception results in acceptance, the picture changes significantly. As the cases described in section 3 showed, one does not necessarily speak against an acceptance by showing that it goes against the evidence. This is because since acceptance 'aims at utility' it is subject to different evaluative norms compared to belief. Thus someone taking the revisionist option will owe us an explanation of what the irrationality of self-deception consists in, assuming she is not inclined to also

[10] Funkhouser treats this as part of his 'Minimal Conception' of self-deception in his recent book on the topic (2019, p. 53), and the word 'irrationality' appears in the titles of many books and articles on self-deception.

revisionistically reject, against near universal opinion, the idea that self-deception is a kind of irrationality.

It therefore appears that it would be much less disruptive to our pre-existing platitudes and considered opinions, both about the nature of self-deception and the nature of belief, to just reject the PHT theory as an account of the dynamics of self-deception. Radically reconceiving the phenomenon to accommodate an explanation of it would be like cutting the foot to fit the shoe. This rejection is made feasible by the availability of alternative explanatory approaches. I will not develop an alternative account of the dynamics of self-deception here, as I have done this elsewhere (Lynch, 2017; 2020).

7. The explanatory relevance of doxastic error-costs for self-deception

Are we being too quick to dismiss the explanatory relevance of doxastic error-costs for self-deception? I would like to suggest that there might be a place for them in the explanation of some cases of self-deception, though the form of the explanation is very different from what PHT theorists have thought it to be.

There are certain cases of twisted self-deception where the belief formation process seems reflex-like (which would make them more akin to cases of wishful thinking than self-deception perhaps). Consider, for instance, a hypochondriac who jumps to the worst conclusion after observing some innocuous symptoms. This phenomenon can perhaps be seen as continuous with something that is widespread in the animal kingdom. Many animals, like birds and house cats, are very easily panicked and flee at the slightest commotion. Fearful creatures that they are, they interpret anything that seems amiss as a potential threat. Similarly, when some people are very worried or fearful of something, they are anxiously on the lookout for anything that might suggest it. A person who is very fearful about having a disease can be in a frame of mind where he is highly sensitive to, and on the lookout for, any indications of that, and is disposed to interpret any ambiguous datum as an ominous sign.

Doxastic error-costs can be relevant to explaining these cases, but we needn't think of them as featuring in a psychological explanation. That is to say, hypochondriacs need not consider or 'compute' the costs of falsely believing certain things *any more than animals do* when they fearfully flee from some disturbance.

Rather, it is simply that when one strongly fears some prospect one is hypersensitive to potential evidence for it and is thus poised to believe it. Where doxastic error-costs enter the picture is in explaining how we evolved to have these sorts of minds. The idea here is that creatures with this mental facility were 'selected' because they tended to avoid the costs of falsely believing certain things. Animals that fled on the slightest provocation tended to avoid the costs of falsely believing that there are no predators around, while the less skittish ones tended to suffer these costs and not reproduce. But this does not imply that they consider or figure out, consciously or unconsciously, the costs of falsely believing that. They just have a mental disposition that helps them avoid these costs and that was selected for that reason, a disposition that is adaptive as opposed to rational. There is no need to 'psychologize' this evolutionary explanation. This kind of perspective is found, for instance, in Öhman and Mineka's theory of an evolved fear module:

> … in fearful circumstances in which rapid defense recruitment is called for, it would have been counterproductive to design the system for requiring a complete cognitive analysis of the situation before defense was activated. Rather the defense response should be automatically activated after only minimal analysis of the stimulus. Clearly, in this type of situation, false negatives (i.e., failure to elicit the defense response in a dangerous situation) would be more evolutionarily costly than false positives (i.e., elicitation of a response to a stimulus that turned out to be harmless …) (2001, p. 502).

So the calculative task of comparing the costs of false negatives with the costs of false positives can be shifted from the subject's mind to the evolutionary process: 'we are designed for relying on the evolutionary wisdom distilled in the fear module rather than for trusting our own thinking' (Ibid., p. 506).

Such an explanation is different from the PHT theory because it rejects the idea that there is any 'computation' (Friedrich, 1993, p. 317) of error-costs taking place. Again, this is best seen from considering non-human animals. When a stray cat flees from an approaching stranger, nobody would suggest that this behaviour proceeded from a judgment, at any level, that the costs of falsely assuming that it is a friend are higher than the costs of falsely assuming that it is a foe. This judgment would involve thoughts with a complex if-then structure and would be 'a significant

cognitive achievement' (Ichikawa *et al*, 2012, p. 341). Positing such cognitive ability in cats would be gratuitous when a primal fear response to unusual stimuli, evolved due to past doxastic error-costs, can explain the behavior. But the cat's behavior might nevertheless roughly match what it *would* do *if* it made such a judgment, though the relatively rigid nature of the mechanism would produce more frequent though harmless false positives.

This is not to suggest that all twisted self-deception could be explained with more simple psychological mechanisms like this, and it is likely that other cases of it are psychologically more complicated. Most philosophers who have set about trying to explain twisted self-deception have done so by positing *ulterior motives* somehow causing the unwelcome belief. For instance, a man believes (or accepts?) that his wife is being unfaithful so that he can take steps to eliminate a rival (Barnes, 1997, p. 44), or so as to avoid disappointment and shock from discovering real infidelity in the future (Nelkin, 2002, pp. 395–396), or someone with a fear of failure believes that she is academically untalented to justify not trying (Funkhouser, 2019).[11] I only wish to point out that doxastic error-costs can be explanatorily relevant to at least some cases of self-deception without our having to suppose that the subject cognizes these error-costs.[12]

Acknowledgments

I would like to thank Davide Fassio, Jie Gao, Melanie Sarzano, Ema Sullivan-Bissett and Zhu Xu, as well as my students at a reading group at Huaqiao University, the audience at a talk at Shanghai Normal University in 2019, and the reviewers of this journal, for helpful comments on this material. Thanks also to Breege Lynch for help with proofreading.

[11] I will not try to come to an understanding of these cases here, but given that there is some prudential or pragmatic ulterior motive involved it is possible that these subjects accept rather than believe these propositions, in which case it would be open to question whether these cases are examples of self-deception.

[12] I would like to thank Davide Fassio, Jie Gao, Melanie Sarzano, Ema Sullivan-Bissett and Zhu Xu, as well as my students at a reading group at Huaqiao University, the audience at a talk at Shanghai Normal University in 2019, and the reviewers of this journal, for helpful comments on this material. Thanks also to Breege Lynch for help with proofreading.

References

Annette Barnes, *Seeing Through Self-Deception* (Cambridge: Cambridge University Press, 1997).

Selim Berker, 'A combinatorial argument against practical reasons for belief', *Analytic Philosophy*, 59 (2018), 427–270.

Michael Bratman, 'Practical reasoning and acceptance in a context', *Mind*, 101 (1992), 1–15.

L. Jonathan Cohen, *An Essay on Belief and Acceptance* (Oxford: Clarendon Press, 1992).

L. Jonathan Cohen, 'Belief and acceptance', *Mind*, 98 (1989), 367–389.

Wayne Davis, 'Knowledge, acceptance, and belief', *The Southern Journal of Philosophy*, 26 (1988), 169–178.

Keith DeRose, 'Contextualism and knowledge attributions', *Philosophy and Phenomenological Research*, 52 (1992), 913–929.

Pascal Engel, 'Believing, holding true, and accepting', *Philosophical Explorations* 1 (1998), 140–151.

Davide Fassio, 'The aim of belief', *Internet Encyclopedia of Philosophy* (2015), https://iep.utm.edu/beli-aim/

Keith Frankish, *Mind and Supermind* (Cambridge: Cambridge University Press, 2004).

James Friedrich, 'Primary error detection and minimization (PEDMIN) strategies in social cognition: a reinterpretation of confirmation bias phenomena', *Psychological Review*, 100 (1993), 298–319.

Eric Funkhouser, *Self-Deception* (Abingdon and New York: Routledge, 2019).

Jie Gao, 'Self-deception and pragmatic encroachment: A dilemma for epistemic rationality', *Ratio*, 34 (2021), 20–32.

Jonathan Ichikawa, Benjamin Jarvis and Katherine Rubin, 'Pragmatic encroachment and belief-desire psychology', *Analytic Philosophy*, 53 (2012), 327–343.

Elizabeth Jackson, 'How belief-credence dualism explains away pragmatic encroachment', *The Philosophical Quarterly*, 69 (2019), 511–533.

Kevin Lynch, 'Review of *Self-Deception*, by Eric Funkhouser', *Philosophy*, 95 (2020), 147-151.

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

Alfred Mele, *Self-Deception Unmasked* (Princeton; Oxford: Princeton University Press, 2001).

Justin McBrayer, 'Review of Aaron Rizzieri (ed.), Pragmatic Encroachment, Religious Belief, and Practice', *Notre Dame Philosophical Reviews* (2014), https://ndpr.nd.edu/reviews/pragmatic-encroachment-religious-belief-and-practice/

Dana Nelkin, 'Self-deception, motivation, and the desire to believe', *Pacific Philosophical Quarterly*, 83 (2002), 384–406.

Arne Öhman and Susan Mineka, 'Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning', *Psychological Review*, 108 (2001), 483–522.

Melanie Sarzano, 'Costly false beliefs: What self-deception and pragmatic encroachment can tell us about the rationality of beliefs', *Les Ateliers de l'ethique/The Ethnics Forum*, 13 (2018), 95–118.

Dion Scott-Kakures, 'At "Permanent Risk": Reasoning and Self-Knowledge in Self-Deception', *Philosophy and Phenomenological Research* 65 (2002), 577–603.

Dion Scott-Kakures, 'Motivated believing: wishful and unwelcome', *Noûs* 34 (2000), 348–375.

Nishi Shah and J. David Velleman, 'Doxastic deliberation', *The Philosophical Review*, 114 (2005), 497–534.

Mark Schroeder, 'Stakes, withholding, and pragmatic encroachment on knowledge', *Philosophical Studies*, 160 (2012), 265–285.

John Skorupski, 2009. 'The unity and diversity of reasons', in S. Roberston (ed.), *Spheres of Reason: New Essays in the Philosophy of Normativity*. Oxford: Oxford University Press, pp. 113–139.

Ema Sullivan-Bissett, 'Aims and exclusivity', *European Journal of Philosophy*, 25 (2017), 721–731.

Yaacov Trope, and Akiva Liberman, 1996. 'Social hypothesis testing: cognitive and motivational mechanisms', in E. T. Higgins and A. W. Kruglanski (eds.), *Social Psychology: Handbook of Basic Principles*, (New York: Guilford, 1996) 239–270.

Yaacov Trope, Benjamin Gervey, and Nira Liberman 1997. 'Wishful thinking from a pragmatic hypothesis testing perspective', in M. S. Myslobodsky, (ed.), *The Mythomanias: The Nature of Deception and Self-Deception*, (New Jersey: Lawrence Erlbaum Associates) 105–131.

Bas van Fraassen, *The Scientific Image* (Oxford: Clarendon Press, 1980).

Edna Ullman-Margalit, 1983. ‘On presumption’, *The Journal of Philosophy*, 80 (1983), 143–163.

Ralph Wedgwood, ‘The aim of belief’, *Philosophical Perspectives*, 16 (2002), 267–297.

Alex Worsnip, ‘Can pragmatists be moderate?’, *Philosophy and Phenomenological Research*, 102 (2021), 531–558.