(This is a penultimate version. The final version is due to be published in *Ratio*. Please refer to the published version.)

Self-deception and pragmatic encroachment: a dilemma for epistemic rationality

Jie Gao
Zhejiang University

Abstract

Self-deception is typically considered epistemically irrational, for it involves holding certain doxastic attitudes against strong counter-evidence. Pragmatic encroachment about epistemic rationality says that whether it is epistemically rational to believe, withhold belief or disbelieve something can depend on perceived practical factors of one's situation. In this paper I argue that some cases of self-deception satisfy what pragmatic encroachment considers sufficient conditions for epistemic rationality. As a result, we face the following dilemma: either we revise the received view about self-deception or we deny pragmatic encroachment on epistemic rationality. I suggest that the dilemma can be solved if we pay close attention to the distinction between ideal and bounded rationality. I argue that the problematic cases fail to meet standards of ideal rationality but exemplify bounded rationality. The solution preserves pragmatic encroachment on bounded rationality, but denies it on ideal rationality.
Keywords: self-deception, pragmatic encroachment, epistemic rationality

0. Introduction

Can self-deception be epistemically rational? The orthodoxy in the psychological and philosophical literature answers this question negatively. Being epistemically rational requires having sufficient evidential support for one's doxastic attitude, and self-deception involves believing against the evidence. Recently, however, the orthodox views of epistemic rationality have been challenged. According to a popular view in contemporary epistemology, *pragmatic encroachment on epistemic rationality*, whether someone has good reasons for her belief depends on a balance between evidence and practical considerations. According to this view, whether it is epistemically rational to believe, withhold belief or disbelieve something can depend on perceived practical factors of one's situation.

The main aim of this paper is to argue that some cases of self-deception satisfy what pragmatic encroachment considers sufficient conditions for epistemic rationality. The result is a dilemma about epistemic rationality: either one accepts that doxastic attitudes in these cases are epistemically rational, and thus is committed to reject the received view that self-deception is always irrational, or one denies that such cases are instances of epistemic rationality, and is thereby forced

to abandon pragmatic encroachment. A further aim of the paper is to suggest a possible solution to this dilemma: I argue that the dilemma can be solved if we pay close attention to the distinction between ideal and bounded rationality. Roughly, the idea is that we should accept that the problematic self-deception cases are epistemically rational given standards of bounded rationality, but not according to standards of ideal rationality. The proposed solution contributes to clarifying in what sense and under which conditions self-deception could be considered rational, and enriches our understanding of pragmatic encroachment.

The paper is structured as follows. In §1 and §2 I introduce respectively self-deception and pragmatic encroachment on epistemic rationality. In §3 I show that some self-deception cases can be classified by pragmatic encroachment as instances of epistemic rationality. These cases reveal the dilemma mentioned above. In §4, I address a possible worry about the existence of the alleged dilemma. In §5, I propose my solution to the dilemma.

1. Self-deception

Marco has received multiple reports from a schoolteacher about his son's insolent and aggressive behaviours towards his classmates. He has already witnessed several instances of such misbehaviour himself. But since Marco has a strong desire that his son be friendly and nice with people and would feel deep shame if this were not the case, he continues believing that his son is a friendly and nice child. Cases like this, where a subject succeeds in maintaining a welcome belief that *p* which has a high subjective importance (such as beliefs related to self-esteem) but is ill-supported by evidence, are paradigmatic instances of *straight self-deception*. Some philosophers have argued that such self-deceptive beliefs have the function of fulfilling the subject's desire for the truth of *p*, quelling her fears or reducing her anxiety in the face of counterevidence to *p* (Barnes 1997; Johnston 1988; Mele 2001).

By contrast, so-called *twisted self-deception* involves believing an unwelcome proposition against persuasive evidence supporting a welcome result. For example, a jealous husband believes on scarce evidence that his wife is having an affair, something that he would like not to be true. Or, someone on the way to work comes to believe that the stove burner at home is still on just because she couldn't properly recall having turned it off and fears the house will burn down.[1]

Although the primary products of self-deception are usually taken to be beliefs, it is reasonable to extend the scope of self-deception to other binary doxastic attitudes, and in particular to withholding belief or suspension.[2] Differently from self-deceived belief, self-deceived withholding

[1] It's disputable whether twisted self-deception also serves the function of reducing anxiety like straight self-deception does. For a defence of an anxiety reduction view see Barnes (1997). See Scott-Kakures (2001) for a criticism of Barnes's view.

[2] It is worth acknowledging that the contemporary literature on self-deception mainly focuses on the attitude of belief. An extension of self-deception to withholding has been recognised by a few authors so far. In particular, Sarzano (2018) explicitly endorses the view that self-deception can be extended to withholding. Funkhouser (2012) provides a scheme generalizing self-deception to a wider variety of attitudes, states and actions that is also applicable to cases of withholding. According to his scheme, conditions that are characteristic of self-deception with respect to some psychological state or behaviour X include: 1) *S* is motivated to X, 2) *S* is in a state that directly conflicts with X, 3) *S* does not have adequate reason to X, 4) *S*

consists of a failure to form a belief in a certain proposition despite very good evidence in its support and withholding belief instead. The extension of self-deception from belief to withholding fits well with the widely held view that withholding is a *sui generis* doxastic attitude on a par with belief and disbelief (Friedman 2013). Like other doxastic attitudes, withholding arguably is also a truth-directed attitude, representing a subject's committed neutrality or indecision with respect to the truth of some proposition. Moreover, withholding seems to be governed by evidential norms like other doxastic attitudes.[3] Since a core feature of self-deception is the violation of evidential norms governing doxastic attitudes, an extension of its scope to withholding seems to be a quite natural and reasonable step. Moreover, as long as withholding can possibly be brought about by the same mechanisms responsible for self-deceptive beliefs, this extension is also compatible with the main accounts of self-deception available in the literature (cf. Barnes 1997; Mele 2001).[4]

In this paper, I remain neutral on the specific type of mechanism that generates self-deception. I only assume that the products of self-deception can be beliefs and withholding.[5] Self-deceptive withholding is particularly important for the main purpose of this paper. As I will explain in the following sections, some paradigmatic cases in which the verdicts of the orthodox view about self-deception and pragmatic encroachment conflict concern agents who withhold belief in self-deceptive ways.

---

employs some deceptive strategies, often involving perversions of rationality, to further X, and 5) *S* has some success in furthering X. This account can be easily applied to cases of withholding. More specifically, *S* is self-deceived in withholding belief with respect to *p* when the following conditions are satisfied: 1) *S* is motivated to withhold belief with respect to *p*, 2) *S* is in a state that conflicts with withholding belief with respect to *p* (i.e. belief that *p*, disbelief that *p*, or absence of opinion), 3) *S*'s total evidence strongly supports *p*/~*p*, 4) *S* employs some deceptive strategies to further withholding belief with respect to *p*, 5) *S* withholds belief with respect to *p*.

[3] As the formation of a belief is the rational response to sufficient evidence for *p*, and disbelief is the rational response to sufficient evidence that ~*p*, it has been argued that withholding is the rational response to insufficient evidence for both *p* and ~*p* (e.g., Sylvan 2016, §3, Archer 2017 and McGrath forthcoming).

[4] Two main accounts of self-deception are intentionalism and motivationalism. The intentionalist account models self-deception on interpersonal deception. According to this account, self-deceived subjects intentionally get themselves to believe *p*, all the while knowing or believing not-*p* (Davidson 1985; Sorensen 1985; Pears 1986; Rorty 1988). Similarly, we can conceive of self-deceived subjects who intentionally get themselves to withhold judgment while knowing or believing *p*. Alternatively, according to the motivationalist account, self-deceptive beliefs are a species of motivationally biased belief, i.e. the result of a subject's motivational state (such as a desire or an emotion) that biases the assessment/estimate of evidence (Johnston 1988; Barnes 1997; Mele 1997; 1999; 2001; Nelkin 2002; Funkhouser 2005; Scott-Kakures 2002; 2012). If suspension is generated in a motivationally biased way and plays a functional role similar to that of self-deceptive beliefs, such as reducing anxiety, there is no reason why we shouldn't also classify such withholding cases as instances of self-deception as well.

[5] It is also worth observing that there have been various accounts of self-deception according to which the products of self-deception are non-doxastic attitudes, such as sincere avowal or a disposition to avow *p* (Audi 1982), pretense (Gendler 2007), imaginations or fantasies that directly express the self-deceiver's wishes, fears, hopes and the like (Lazar 1999) and some intermediate state between belief and desire (Egan 2009). I recognise that the dilemma I discuss below cannot be constructed if the product of self-deception is non-doxastic. Hence the claim in this paper can be understood as conditional: if the product of self-deception can be doxastic, then the dilemma arises.

2. Pragmatic encroachment on epistemic rationality

According to an influential approach in contemporary epistemology, there are close connections between epistemic notions such as knowledge and justification and practical factors. Proponents of *pragmatic encroachment on knowledge* argue that whether someone knows something partially depends on (actual or perceived) practical factors of her situation (e.g. Hawthorne 2004; Stanley 2005; Fantl and McGrath 2009). Pragmatic encroachment on epistemic rationality is closely related to pragmatic encroachment on knowledge and is supposed to provide straightforward support to it.

One prominent argument for *pragmatic encroachment on epistemic rationality* (henceforth, PEER) appeals to a close connection between knowledge-level justification and action or practical reasoning. For example, Jeremy Fantl and Matthew McGrath (2009) defend the following principle concerning knowledge-level justification:

> (JJ) You are justified in believing that *p* iff *p* is warranted enough to justify you in ϕ-ing, for any ϕ.

Naturally, whether one can reasonably take *p* as a premise in practical reasoning is constrained by practical factors specific to the circumstances. For example, a certain epistemic position with respect to *p* may be good enough for one to act on *p* in a low-stakes situation where it is not important to be right about *p,* but not good enough in a high-stakes situation where costs of error are very high. If, as (JJ) suggests, whether one is justified in believing *p* depends on whether one is warranted enough to act on *p*, then practical factors such as stakes are relevant in determining whether one is justified in believing *p*.

Another argument for pragmatic encroachment on epistemic rationality has been put forward by Schroeder (2012). Schroeder moves from the intuitive claim that it is epistemically rational for someone to believe *p* just in case her evidence adequately supports *p*. According to Schroeder, 'adequate support' doesn't merely depend on how much evidence one has and how good it is, but is also partly determined by pragmatic considerations. Schroeder suggests that rationality conditions for belief should be modelled on those for rational action. The latter can be expressed by the following principle:

> (General Sufficiency)
> It is rational for *S* to do *A* just in case *S* has at least as much reason to do *A* as in favour of any of the alternatives to doing *A*.[6]

If we want to build up an account of epistemic rationality modelled on the above principle, we should first identify what the alternatives to believing *p* are. Following Harman, Schroeder holds that the alternatives do not only include believing ~*p* but also withholding belief with respect to *p*. Hence from (General Sufficiency) we can derive the following principle about epistemic rationality:

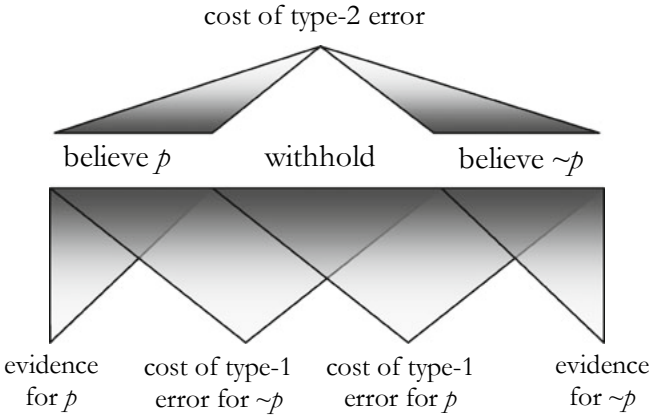
> (Belief Sufficiency)

[6] This is derived from Harman (2002)'s original principle which says that it is rational for *S* to do A just in case *S* has at least as much reason to do *A* as to not do *A*. Schroeder takes (General Sufficiency) to be a natural generalization of the original principle from a two-option case to a multiple-option case.

> It is epistemically rational for *S* to believe *p* just in case *S* has at least as much epistemic reason to believe *p* as to believe ~*p* and *S* has at least as much epistemic reason to believe *p* as to withhold with respect to *p*.

Now, the central issue is what counts as reasons to withhold. According to Schroeder, such reasons cannot be evidence. This is because evidence either supports or disfavours a certain proposition. But evidence supporting *p* is reason to believe *p*, and evidence supporting ~*p* is reason to believe ~*p*. Reasons to withhold then have to be something other than evidence. Schroeder identifies such reasons with the *costs of error*. These come in two types. One type—or *type-1 error*, henceforth $Err1_p$—is associated with having a false belief that *p*, i.e. the cost of being wrong about whether *p* given that one believes *p* and acts on this belief as if *p* were the case. This type of cost constitutes reason not to believe *p*—i.e. either to withhold or to believe ~*p*. The other type of cost—*type-2 error*, henceforth Err2—is associated with missing out on having a true belief, i.e. the cost of suspending judgment and the benefit of having made up your mind. This type of cost constitutes reason to not withhold—to either believe *p* or believe ~*p*.

Schroeder considers costs of error as epistemic reasons on a par with evidence.[7] On this view, the epistemically rational doxastic attitude is a result of a balance between evidential and practical considerations. This balance is illustrated in the following picture, in which shaded triangles represent the support of each kind of reason (Schroeder 2012, p. 279):



In this model, there are two ways in which evidence for *p* (henceforth Ev_p) is better than evidence for ~*p* (henceforth $Ev_{\sim p}$) but it is not rational to believe *p*.[8] First, it would be rational to withhold belief with respect to *p* rather than to believe *p* when the costs of having a false belief that *p* outweigh the costs of missing out on having a true belief by a sufficient amount—enough to outweigh the evidence for *p*:

$$\text{Withholding is more rational: } Err1_p - Err2 > Ev_p$$

[7] Another motivation Schroeder provides for classifying costs of errors as epistemic reasons to withhold is that we directly respond to those costs in a way similar to how we respond to evidence. In this sense, costs of error are different from Pascalian reasons (e.g., rewards to form a certain belief), which do not directly cause our attitudes' formation, but instead require indirect strategies for acquiring doxastic attitudes.

[8] The deductions are omitted for reasons of space. See Schroeder (2012, p. 280) for details.

Second, in cases of forced decision (when Err2 is high), it would be rational to believe $\sim p$ when the costs of having a false belief that p exceed the costs of having a false belief that $\sim p$ to a higher degree than the degree to which the evidence for p outweighs the evidence for $\sim p$:

$$\text{Believing } \sim p \text{ is more rational: } \mathrm{Err1}_{p} - \mathrm{Err1}_{\sim p} > \mathrm{Ev}_{p} - \mathrm{Ev}_{\sim p}$$

The above two types of situation can be instantiated by some high-stakes cases where the costs of having a false belief in an evidentially well-supported proposition are very high. I will call such cases *special high-stakes cases* (henceforth SHS).

Consider two examples of SHS suggested by Schroeder (2012, pp. 281-282):

> (*Nasa Engineering*). Hannah and Sarah are engineers working on the design of NASA's next-generation shuttle, a multi-billion dollar project planned to operate over several decades and ultimately carry hundreds of astronauts into space, where error means death. Currently they are trying to decide which materials to use for an important component, and are investigating two new alloys, to see which will be more appropriate for the component. Citing preliminary research, Sarah notes that the first alloy holds up better under temperatures under 300°, and that most alloys which hold up well under 300° also perform well at shuttle temperatures. Hannah says, 'okay, so the first alloy will hold up better at shuttle temperatures.' In fact Hannah is correct; the first alloy does hold up better at higher temperatures.
>
> (*Game Show*). Hannah and Sarah are playing Go Big or Go Home, a successful game show on daytime television with a B-celebrity host. They have reached the final question, which is: 'will the bank be open tomorrow, on Saturday?'. The possible answers are 'yes' and 'no', and they must answer within the time limit, or they will lose all of their money (they have accumulated a very large sum so far). If they answer and get it right, they double their money, but if they answer 'yes' and get it wrong, they lose all of their money and if they answer 'no' and get it wrong, they keep what they already have. Hannah tells Sarah, 'The answer is 'yes'—I was there three weeks ago on a Saturday morning, and it was open.' In fact, Hannah is correct; the bank will be open tomorrow.

In Nasa Engineering, the cost of falsely believing that the first alloy will hold up at shuttle temperature is super-high but there is no urgency to take an immediate decision. Hence according to Schroeder's theory it is rational for the engineers to withhold belief. In Game Show, presumably the cost of falsely believing that the bank will be open tomorrow exceeds the cost of falsely believing the contrary to a degree that outweighs the evidence for the claim that the bank will be open tomorrow.

Hence according to Schroeder's theory in this circumstance it is rational for the subject to believe against the evidence.[9]

3. A dilemma about epistemic rationality

Having clarified self-deception and PEER, we are now in a position to identify *overlapping cases*, i.e. cases that can be classified as both self-deception and SHS. These are the cases that may pose a dilemma for epistemic rationality—in which attitudes are deemed to be rational according to PEER, but also constitute instances of self-deception, and thus are irrational according to orthodox views. Below I will consider overlapping cases involving, respectively, straight self-deception (§3.1) and twisted self-deception (§3.2).

3.1. Overlapping cases with straight self-deception

Paradigmatic straight self-deception features a strong desire for the truth of a welcome proposition *p*. In these cases, believing the unwelcome but warranted proposition ~*p* is associated with a high emotional cost, i.e. the psychological discomfort which accompanies believing ~*p*. Notice that the emotional cost involved in straight self-deception is about the believing status itself regardless of whether *p* is true or not. Although this emotional cost can in the end lead to practical losses for the subject, this cost is not about the practical consequence of acting on *p* and being wrong. I call such type of cost the *cost of believing*. This should not be confused with the cost of falsely believing, or cost of error (type-1 error).

An overlapping case of straight self-deception and SHS should satisfy the following conditions:

i) a high cost of believing *p*
ii) *S* has good overall evidence for *p*
iii) *S* withholds belief with respect to *p* or believes ~*p*

} features of straight self-deception

iv) $Err1_p - Err2 > Ev_p$, and/or
v) $Err1_p - Err1_{\sim p} > Ev_p - Ev_{\sim p}$

} features of SHS

Consider the following case:

> (*Affair*) Allison has quite good evidence that her husband is having an affair with another woman. At the same time, Allison herself is also having an affair and she knows that her husband has no idea about that. Allison is a very successful business woman. By contrast, her husband is just an ordinary staffer at a bureau. According to the prenuptial agreement, the divorce procedure includes a fidelity investigation of both parties. If Allison has an affair while her husband does not, then her husband is entitled to receive half of Allison's property, but if both have affairs, then her husband will not receive anything. Believing that her husband is having an affair would deeply hurt Allison's self-esteem. In that case she will sue

[9] It is worth noticing here that Schroeder (2012) admits that the believing-the-contrary case is more controversial than the withholding case.

for divorce, which will trigger investigations of both her and her husband. But that also means that she cannot hide her own affair anymore.

In the above case, it would be very painful for Allison to believe that her husband is having an affair (henceforth *a*). The cost of believing *a* is high: condition (i) is satisfied. Allison has good evidence for *a*, so condition (ii) is also satisfied. Assume that condition (iii) is also met: Allison withholds belief with respect to *a* or believes ~*a*. So the attitude constitutes an instance of self-deception.[10]

Presumably, condition (iv) is fulfilled. The cost of falsely believing *a* ($Err1_a$) is very high: acting on *a* and being wrong would lead Allison to lose half of her property. This is not a situation of forced decision, so the cost of missing out on the truth (Err2) is negligible. Thus the difference between the two types of cost could be so large that it outweighs the total evidence for *a*. According to PEER, it would then be rational for Allison to withhold belief with respect to *a*.

Concerning condition (v), we may assume that the cost of falsely believing ~*a* ($Err1_{\sim a}$) is negligible, so it is very likely that $Err1_a$ exceeds $Err1_{\sim a}$ to a degree higher than Ev_a outweighs $Ev_{\sim a}$. Again, according to PEER it would be rational for the subject to believe ~*a*.

In conclusion, on the one hand, it seems that Allison's resultant doxastic attitude can be considered the product of self-deception. Hence according to the orthodox view about self-deception, Allison's attitude cannot be rational. On the other hand, according to PEER, Allison's resultant doxastic attitude may well be rational. Here is the dilemma. The same attitude cannot be both rational and irrational. We must either revise the orthodox view about self-deception or give up PEER.

3.2. Overlapping cases with twisted self-deception

Differently from straight self-deception, instead of bringing more relief to the subject, twisted self-deception brings more torment. The cost of having the resultant doxastic attitude, believing the unwarranted proposition ~*p* or withholding belief, is rather high. Thus, an overlapping case of twisted self-deception and SHS should satisfy the following conditions:

i) a high cost of withholding or believing ~*p*
ii) *S* has good overall evidence for *p*
iii) *S* withholds belief with respect to *p* or believes ~*p*
} features of twisted self-deception

iv) $Err1_p - Err2 > Ev_p$, or
v) $Err1_p - Err1_{\sim p} > Ev_p - Ev_{\sim p}$
} features of SHS

Consider the case of Suzuki, a jealous husband who has quite good evidence that his wife is faithful (henceforth *f*)—say, Ev_f = testimony from several reliable mutual friends, $Ev_{\sim f}$ = small changes in his spouse's behavior statistically associated to people having an affair—so that condition (ii) is satisfied. Believing that his wife is unfaithful, or even withholding belief, will only bring about more jealousy,

[10] A further possible requirement for self-deception which may be satisfied in this case is that condition (iii) obtains because of condition (i), i.e., withholding belief with respect to *p* or believing ~*p* because of the costs of believing *p*. I will say more about this requirement in the next section where I will discuss a possible objection to the existence of the dilemma.

fear or anxiety; so condition (i) is satisfied. Assume that condition (iii) is also met: Suzuki withholds belief or believes ~*f*. We then have a twisted self-deception case.

Let's now consider condition (v), i.e. $Err1_f - Err1_{\sim f} > Ev_f - Ev_{\sim f}$. Suppose that if Suzuki believes *f,* he would not change his attitude towards his wife. In that case, the cost of falsely believing *f* is letting the relationship between his wife and her (presumed) lover develop, with serious chances that his wife will in the end decide to divorce. Suppose also that divorce would be the most devastating thing that could happen to Suzuki. Moreover, he is pretty sure that if that were to happen he couldn't preserve his mental health, becoming a serious danger for himself and other people. So the cost of falsely believing *f* (i.e. $Err1_f$) is extremely high for Suzuki. On the other hand, by believing ~*f,* Suzuki is aware that he would take positive initiatives to strengthen his relationship with his wife, engaging in actions that would enhance the couple's intimacy and mutual trust, ultimately ensuring that his wife will remain faithful in future. For example, Suzuki may spend more time with his wife, organize interesting activities together, buy gifts for her and so on. We can assume that the costs of doing all these things (i.e. $Err1_{\sim f}$) are negligible for Suzuki. Thus, the difference between the two costs of error could be so large that it outweighs the difference in degree between the evidence for *f* and for ~*f*—evidence that we assumed to be quite good, but not decisive. According to PEER, it may then be rational for the subject to believe ~*f*.[11]

Moreover, given further details, this case could be a situation of non-forced decision. For example we can imagine that the subject is sure that his wife will not leave him any time soon given her commitment to raise their child together.[12] If so, the cost of withholding (Err2) would be negligible. Thus condition (iv), i.e. $Err1_f - Err2 > Ev_f$, would be easily met. According to PEER, it would then be rational for the subject to withhold belief with respect to *f*.

No matter how we interpret the case—as a forced decision in which (v) is satisfied, or a non-forced decision in which (iv) is—this instantiates our dilemma: either we accept PEER and we say that the husband is epistemically rational to withhold belief or believe ~*f*, or we follow orthodox views about self-deception and deem these attitudes as irrational.

We have just seen that paradigmatic twisted self-deception cases can constitute instances of SHS. Conversely, typical high-stakes cases discussed in the pragmatic encroachment literature can also be framed as twisted self-deception cases. Consider the following modified high-stakes scenario, inspired by a similar one in Ross and Schroeder (2015, p. 216):

> (*Sandwich*) Hannah prepares two sandwiches, an almond butter sandwich and a tuna sandwich, and places them in the refrigerator. She tells Sarah about that.

---

[11] Assuming that evidence and the costs of error are somehow commensurable, in Suzuki's case we could interpret the assignment of values as follows: $Err1_f = 10$, $Err1_{\sim f} = 0.5$, $Ev_f = 8$, $Ev_{\sim f} = 1$. These assignments would satisfy condition (v). While these assignments are arbitrary, I think that they could somewhat reflect the values of costs of error and balance of evidence in the case. However, if someone disagrees with my assessments, she is free to modify or add further details to the case in ways that would go in the direction of validating the condition. This could be done in three ways: i) by increasing the cost of falsely believing that *f* (e.g., assuming catastrophic consequences of falsely believing that *f*), ii) by lowering the result of $(Ev_f - Ev_{\sim f})$, either slightly lowering the strength of evidence for *f* or slightly increasing the evidence for ~*f*, and iii) by further lowering the cost of falsely believing that ~*f*.

[12] By contrast, it is natural to interpret the stove burner case as a case of forced decision.

> Hannah then leaves just as Sarah's nephew Algernon arrives for lunch. Algernon has a severe peanut allergy. He asks Sarah for a sandwich. Sarah knows that he would very much prefer the almond butter sandwich to the tuna sandwich. Both these sandwiches would be harmless. She also knows that the peanut butter sandwich would be fatal to Algernon and that Hannah sometimes makes peanut butter sandwiches instead of almond butter ones. Sarah cannot tell, by visual inspection, the almond butter sandwich from the peanut butter sandwich.

In this case, the cost of falsely believing that Hannah has made an almond butter sandwich instead of a peanut butter one is very high: a possible mistake could lead to her nephew's death. Again, according to PEER, it is rational for Sarah to either withhold judgment or believe that Sarah has made a peanut butter sandwich.

The above case can be classified as an instance of twisted self-deception. That Hannah has made an almond butter sandwich is a desirable truth, for Sarah would then be able to treat her nephew with the sandwich he prefers. Withholding belief or believing that Hannah has made a peanut butter sandwich will make Sarah feel unhappy because she would not be able to satisfy her nephew's desire. So condition (i) for twisted self-deception is satisfied. Condition (ii) is apparently met because Hannah just told Sarah about the types of sandwich she made. Assume that condition (iii) is also met, and we have a case of twisted self-deception. Cases such as (Sandwich) are commonly used to illustrate PEER, and are supposed to convey the intuition that Sarah would be rational to withhold belief or believe that the sandwich is a peanut butter one. However, according to the orthodox view, as a product of self-deception Sarah's attitudes should be considered irrational. Note that cases such as (Sandwich) are actually very similar to others such as the stove burner case mentioned in §1, in which a person disbelieves that he has turned off the stove burner despite having done it just ten minutes ago. However, the stove burner case is standardly used to illustrate unreasonable self-deception.

Apparently, there seem to be only two ways to solve this dilemma: either we revise orthodox views about self-deception, or we deny PEER. Given that PEER is supposed to explain pragmatic encroachment on knowledge, taking the second horn would naturally lead to a denial of the latter as well. Opponents of pragmatic encroachment will consider this puzzle as a further reason to reject that view—thus endorsing the second horn of the dilemma. While I do not have much stakes in the pragmatic encroachment debate, in this paper I would like to suggest a different solution to the dilemma, one that avoids either horn. However, before discussing my favorite solution, I should first address a potential worry about the existence of the dilemma.

4. A potential worry and response

Melanie Sarzano (2018) also notices a similarity between self-deception cases and restricted high-stakes cases. She points out that specific psychological mechanisms supposed to generate self-deception are also very likely to be employed by high-stakes subjects. But she is inclined to deny that this could constitute genuine dilemmas. She suggests that the type of cost bringing about self-deceived beliefs (i.e. the cost of believing) in straight self-deception cases is different from the type

of cost featured in the belief-forming mechanism in SHS cases (i.e. costs of error). This difference would provide a principled way to distinguish straight self-deception cases from SHS cases. One could exploit a similar strategy to solve our dilemma. The idea is to separate the respective jurisdictions of PEER and self-deception: self-deception would only occur in cases where the resultant attitude is caused by costs of believing, while PEER would only apply to cases where the resultant attitude is caused by costs of error. If in the relevant cases the resultant doxastic attitudes are caused by costs of believing, they would count as instances of irrational self-deception. Conversely, if the resultant attitudes are caused by the costs of error, they would not be instances of self-deception and could be considered rational as predicted by PEER. This strategy would exclude the possibility of overlapping cases, and thus would avoid the dilemma.

Even if we agree with Sarzano's diagnosis about straight self-deception cases, it is worth observing that there are other overlapping cases that can resist a similar diagnosis. So the dilemma still holds for these cases. In particular, the above diagnosis does not apply to the twisted self-deception cases considered in the previous section. As several self-deception theorists agree, it is implausible that the costs of believing play any role in causing twisted self-deception. After all, the resultant attitudes in twisted self-deception cases only bring immediate psychological discomfort. If there is any type of cost operative in bringing about twisted self-deception, it has to be the cost of error (Scott-Kakures 2000, Mele 2009). Consider the jealous husband case again. It is natural to interpret Suzuki's self-deception as a consequence of being overwhelmed by the costs of being wrong about whether his wife is faithful. Thus, we cannot clearly distinguish the types of costs featured in the mechanism involved in twisted self-deception and in SHS. While self-deception theorists take costs of error as causing irrationality, proponents of pragmatic encroachment take them as determinants of epistemic rationality.

I also disagree with Sarzano that her diagnosis is sufficient to avoid any sort of dilemma in overlapping cases involving straight self-deception. Let's distinguish two senses of epistemic rationality: one concerns propositional justification, i.e. whether one *has* good reasons for one's doxastic attitude; the other concerns doxastic justification, i.e. whether one's doxastic attitude *is based* on good reasons. When we assess an attitude's propositional justification, we shouldn't consider on which basis the attitude is formed and retained, that is, the mechanisms of attitude formation and regulation. These mechanisms include whether the attitude is triggered by costs of believing or costs of error. Once we consider such cases from an *ex ante* perspective, and we ask what the subject in this situation should believe, the discriminatory criterion suggested by Sarzano (i.e. the distinction between attitudes' formation influenced by costs of error vs. cost of believing) becomes irrelevant. PEER would claim that in such cases it is rational (propositionally justified) to have the resultant attitude, while the traditional view of self-deception would still deem the relevant attitude as irrational. Therefore, although the relevant cases may not be problematic when we assess the attitudes' doxastic justification, they still constitute a dilemma for propositional justification.

Moreover, we should also consider cases where both costs of believing and costs of error contribute to, or overdetermine, the formation and retention of the resultant doxastic attitude. Sarzano's strategy doesn't apply to such type of case, since the two types of cost are intertwined in, and co-responsible for, the attitude's formation. Such cases could also present a dilemma for the

attitudes' doxastic justification. As long as the costs of believing play a non-idle role in the resultant attitude's formation, no matter whether its role is essential or not, the case involves at least *partial straight self-deception*. According to the orthodox view, the product of partial self-deception cannot be doxastically justified, whereas it is in the spirit of PEER to ascribe some degree of doxastic justification to the relevant doxastic attitudes given that the costs of error significantly contribute to the attitude's formation.

In conclusion, overlapping cases involving twisted self-deception seem clearly to constitute genuine dilemmas. Moreover, I argued that overlapping cases involving straight self-deception pose at least dilemmas for what concerns propositional justification, and a specific range of partial straight self-deception cases may constitute dilemmas for doxastic justification also.

5. A tentative solution to the dilemma

The dilemma presented so far demands either a revision of orthodox views about self-deception or a denial of pragmatic encroachment on epistemic rationality. The second horn seems to be difficult to defend. Taking this horn would not just require a denial of a popular view in contemporary epistemology, but also require holding the very counterintuitive idea that withholding belief with respect to the relevant proposition in SHS cases such as Sandwich is irrational. The first horn, by contrast, is not as radical as it might appear. Since only particular instances of self-deception fall into the relevant category of overlapping cases, we are not forced to a substantial revision of the orthodox view. Rather, the revision would affect only the problematic overlapping cases. Nonetheless, saying that those cases of self-deception are completely epistemically rational would sound to many as far too radical. After all, if the reason why we attribute irrationality to standard self-deception is because it's formed on the basis of inadequate truth-relevant considerations (e.g., insufficient evidence), this diagnosis should carry over to the overlapping cases as well.

A viable solution to the dilemma should be able to explain the intuitive pulls to attribute both rationality and irrationality to the attitudes in the overlapping cases. In this section I aim to provide such a solution. My suggestion relies on the claim that there is an ambiguity in the notion of epistemic rationality involved in the respective judgments. In short, I suggest that we should accept that self-deception in overlapping cases can be epistemically rational, but in a qualified sense. More specifically, I propose that these instances of self-deception are rational given standards of bounded rationality but not according to standards of ideal rationality.[13]

Unbounded, ideal rationality takes truth and accuracy as its only standards, abstracting away from limitations of one's cognitive abilities. Ideal epistemic agents' doxastic attitudes would obey strict principles based on rules of logic and probability theory. In addition, these doxastic attitudes should be completely isolated from influences of non-truth-relevant factors, such as psychological, emotional, practical or environmental factors. Ideally rational attitudes are supposed to be responsive only to truth-relevant factors such as accuracy, evidential support or the attitude's degree of reliability.

[13] For overviews of discussions and relevant literature on different types of rationality, see e.g. Samuels et al. (2004), Hertwig and Pedersen (2016) and Gao (2019).

Self-deception deviates from the requirements of ideal rationality, since self-deceived belief and withholding remarkably fail to be exclusively sensitive to truth-relevant factors. A self-deceived agent either believes, and hence holds a very high credence, in a proposition on dubious evidential grounds, or she assigns a low credence falling short of full belief despite the evidence strongly favouring a given proposition.

However, human beings are not ideal rational agents. In real circumstances, our cognitive performances are bounded by serious physical, ecological and temporal limits. Even though the human mind commits to certain patterns of cognition (such as heuristics, fallacies and biases) that are not recommended by standards of ideal rationality, humans should not be considered systematically and irremediably unreasonable. *Bounded rationality* characterizes the type of rationality typical of normal human beings constrained by limitations of mental and environmental resources.[14] A theory of bounded rationality focuses on the structure of the environment and on the adaptations of one's cognitive system to it. In this picture, a cognitive system should count as manifesting bounded rationality if it is functional in the achievement of epistemic goals such as acquiring truth and avoiding error and is compatible with limits imposed by one's cognitive abilities and the environment. Natural cognitive systems are designed to achieve a proper balance between epistemic goals such as accuracy and reliability and practical purposes and needs.[15]

The sensitivity of doxastic attitudes to practical considerations (i.e. costs of error) as suggested by PEER meets the standards of bounded rationality in at least two respects. First, it allows us to wisely allocate our energy given the demands of specific practical and cognitive tasks, leading us to form settled beliefs on a certain issue when the accuracy of judgment is good enough for a given purpose. Our cognitive limitations do not allow us to carry on perpetual inquiries on a given issue. The search and deliberation must end at some point. But it doesn't seem reasonable to terminate our inquiry and form a settled opinion in an arbitrary way, or consciously leave some relevant evidence out of consideration, on pain of forming judgments based on too shaky and inaccurate grounds. By connecting belief-formation with our specific practical needs, the practical sensitivity of doxastic attitudes predicted by PEER provides a natural mechanism for allocating our cognitive efforts in efficient and acceptably reliable ways.

Second, postponing belief formation and withholding when stakes are high prevents us from acting imprudently. Since normally belief doesn't imply absolute certainty or maximal strength of justification, acting on our fallible beliefs implies some risk of making bad choices. Prudence recommends that we reduce such risk to a minimum when stakes are high, allowing for wide margins of safety from error. The practical sensitivity of doxastic attitudes enables us to postpone belief formation and action until one's epistemic position has been sufficiently strengthened, leading to more cautious and less risky choices.

In sum, the doxastic attitudes' sensitivity to practical factors predicted by PEER helps human beings in better allocating their cognitive efforts and protects them from dangerous mistakes. It thus

[14] According to a famous analogy suggested by Hebert Simon, "Human rational behaviour…is shaped by a scissors whose two blades are the structure of the task environments and the computational capabilities of the actor" (Simon 1990: 7).

[15] See Gao (2019) for why we should classify bounded rationality as a type of epistemic rationality.

seems appropriate to attribute rationality to such sensitivity, even though only in a qualified sense. Our dilemma can be solved by qualifying rationality judgments about self-deception: while self-deception is never ideally rational, in overlapping cases it can be considered boundedly rational. This solution relies on an independently motivated distinction between different types of epistemic rationality and has the merit of providing a simple and straightforward explanation of the seemingly contradictory intuitive pulls to attribute both rationality and irrationality to attitudes in overlapping cases.

Reference

Archer, Sophie. 2017. "Defending Exclusivity." *Philosophy and Phenomenological Research* 94 (2): 326–341.

Audi, R. 1982. "Self-Deception, Action and Will." *Erkenntnis* 18: 133–58.

Barnes, Annette. 1997. *Seeing through Self-Deception*. Cambridge: Cambridge University Press.

Davidson, D. 1985. "Deception and Division." In *Actions and Events*, edited by E. LePore and B. McLaughlin, 79–92. New York: Basil Blackwell.

Egan, A. 2009. "Imagination, Delusion, and Self-Deception." In *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, edited by T. Bayne and J. Fernandez, 263–80. New York: Psychology Press.

Fantl, Jeremy, and Matthew McGrath. 2009. *Knowledge in an Uncertain World*. Oxford: Oxford University Press.

Friedman, Jane. 2013. "Suspended Judgment." *Philosophical Studies* 162 (2): 165–81.

Gao, J. 2019. "Credal Pragmatism." *Philosophical Studies* 176 (2): 1595–1617.

Gendler, T. 2007. "Self-Deception as Pretense." *Philosophical Perspectives* 21: 231–58.

Hawthorne, John. 2004. *Knowledge and Lotteries*. Oxford: Oxford University Press.

Hertwig, Ralph, and Arthur Pedersen. 2016. "Find Foundations for Bounded and Adaptive Rationality." *Minds & Machines* 26: 1–8.

Johnston, M. 1988. "Self-Deception and the Nature of Mind." In *Perspectives on Self-Deception*, edited by Brian McLaughlin and A. O. Rorty. Berkeley: University of California Press.

Lazar, A. 1999. "Deceiving Oneself or Self-Deceived?" *Mind* 108: 263–90.

McGrath, M. forthcoming. "Being Neutral: Agnosticism, Inquiry and the Suspension of Judgment." *Noûs*.

Mele, Alfred. 2001. *Self-Deception Unmasked*. Princeton: Princeton University Press.

Nelkin, Dana. 2002. "Self-Deception, Motivation, and the Desire to Believe." *Pacific Philosophical Quarterly* 83: 384–406.

Pears, D. 1986. "The Goals and Strategies of Self-Deception." In *The Multiple Self*, edited by J. Elster, 59–78. Cambridge: Cambridge University Press.

Rorty, A. O. 1988. "The Deceptive Self: Kiars, Layers, and Lairs." In *Perspectives on Self-Deception*, edited by B. McLaughlin and A. O. Rorty, 11–28. Berkeley: University of California Press.

Samuels, Richard, Stephen Stich, and Luc Faucher. 2004. "Reason and Rationality." In *Handbook of Epistemology*, edited by Ilkka Niiniluoto, Matti Sintonen, and Jan Wolenski, 131–79. Dordrecht: Kluwer Academic Publishers.

Sarzano, Melanie. 2018. "Costly False Beliefs: What Self-Deception and Pragmatic Encroachment Can Tell Us about the Rationality of Beliefs." *The Ethics Forum* 13 (2): 95–118.

Schroeder, Mark. 2012. "Stakes, Withholding, and Pragmatic Encroachment on Knowledge." *Philosophical Studies* 160 (2): 265–85.
Scott-Kakures, D. 2001. "High Anxiety: Barnes on What Moves the Unwelcome Believer." *Philosophical Psychology* 14: 348–75.
———. 2002. "At Permanent Risk: Reasoning and Self-Knowledge in Self-Deception." *Philosophy and Phenomenological Research* 65: 577–603.
———. 2012. "Can You Succedd in Intentionally Deceiving Yourself?" *Humana. Mente Journal of Philosophical Studies* 5 (20): 17–39.
Simon, H. A. 1990. "Invariants of Human Behavior." *Annual Review of Psychology* 41: 1–19.
Sorensen, R. 1985. "Self-Deception and Scattered Events." *Mind* 94: 64–69.
Stanley, Jason. 2005. *Knowledge and Practical Interests*. Oxford: Oxford University Press.
Sylvan, Kurt. 2016. "The Illusion of Discretion." *Synthese* 193 (6): 1635–1665.