SELF-DECEPTION AND EPISTEMIC BLAME

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Zeinab Rabii

Dr. Paul Weirich, Dissertation Supervisor

December 2023
The undersigned, appointed by the dean of the Graduate School, have examined the
dissertation entitled

SELF-DECEPTION AND EPISTEMIC BLAME

presented by Travis Holmes,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

____________________________
Professor Paul Weirich

____________________________
Professor Kenneth Boyce

____________________________
Professor Robert Johnson

____________________________
Professor Randall Westgren
For my parents: Abolfazl Rabii and Zahra Beheshtirooy
ACKNOWLEDGEMENTS

I couldn’t have reached this goal without the help of many people in my life. But I would like to take this opportunity to thank each member of my community and my family.

First, special thanks are owed to Paul Weirich. Words can't adequately express my gratitude for what Dr. Weirich has done for me during these years of graduate school at Missouri. He compassionately showed me how to improve my projects. I have been very lucky to have his continuous support and invaluable advice all the time. Undoubtedly, without his supervising, I could not obtain my academic goals and finish my dissertation.

Second, thanks also go out to Kenneth Boyce who patiently worked with me and read my papers carefully. I am very grateful for Kenny’s insightful comments on different versions of my papers, and also great support and guidance both for my MA thesis and doctoral dissertation.

Thank you also to Robert Johnson who kindly helped me with my dissertation. I have learned a lot from him both as a teaching assistant and a graduate student in his classes.

I will also forever be thankful to Randall Westgren who graciously agreed to serve on this committee. Dr. Wesgren’s words of encouragement motivated me to develop my research projects.

And finally, but perhaps most importantly, I must express my profound gratitude to my family: my father, my mother, Somayeh, Morteza, Ali, Arghavan and Hasti. Thinking of you brightens my life every day. Although I have not seen you for six years, but there has never been a day on which I did not receive your unfailing support and
encouragement. I would like to extend my special thanks to my spouse, Travis Holmes, as well for his great company and support during these years. Without these people, I could not obtain my life dreams.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>ii</td>
</tr>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Paper One</td>
<td>3</td>
</tr>
<tr>
<td>Paper Two</td>
<td>26</td>
</tr>
<tr>
<td>Paper Three</td>
<td>50</td>
</tr>
<tr>
<td>Conclusion</td>
<td>75</td>
</tr>
<tr>
<td>Bibliography</td>
<td>78</td>
</tr>
<tr>
<td>VITA</td>
<td>82</td>
</tr>
</tbody>
</table>
ABSTRACT

This dissertation consists of three papers, and each of the papers has a main question: a) how do we reconcile the knowledge thesis (namely, a view which says that it is impossible for an agent to act intentionally without knowing what she is doing) with self-deception? b) Do self-deceivers really believe a justified unpleasant proposition that not-\( P \), while only accepting a pleasant unjustified proposition that \( P \) without believing it? c) What is the best account of epistemic blame?

In the first chapter of my dissertation, I explain and then criticize Eric Marcus’ (2019) solution for the reconciliation between the knowledge thesis and self-deception. Then, I offer an account of a self-deceiver’s knowledge which shows how self-deception is compatible with the knowledge thesis.

In the second chapter of my dissertation, I lay out different views of the belief-acceptance distinction and show that the views at issue cannot properly solve the static puzzle of self-deception (namely, a puzzle according to which a self-deceiver holds two contradictory beliefs simultaneously), in spite of Erik Funkhouser’s (2019) claim that the belief-acceptance distinction is the best strategy to solve the puzzle.

And the final chapter of my dissertation belongs to epistemic blame. In this chapter, after explaining four different theories of epistemic blame (namely, the relationship-based view, the agency-cultivation account, the desire-based view, and the emotion-based theory), I pick out the emotion-based theory of epistemic blame as the only theory potential to be revised. After revising the theory, I show why this theory has better explanatory power than the other theories of epistemic blame.
INTRODUCTION

In my first paper, I argue for the compatibility of self-deception with the knowledge thesis (KT). According to the Knowledge Thesis, it is impossible for an agent to act intentionally without knowing what he is doing. Eric Marcus (2019) defends this view, and attributes it to Elizabeth Anscombe (2000). Marcus states that self-deceptive action (namely, cases of self-deception in which an agent non-lyingly disavows knowledge of what he is doing intentionally) is among the most compelling counterexamples of KT, and offers an account of self-deceptive action compatible with KT. In the first paper of my dissertation, I aim to show that Marcus’ project fails, since his account of self-deception is problematic. However, this doesn’t mean that there is no way to show the compatibility of self-deception with KT. I argue that the existence of self-deception doesn’t threaten KT by introducing a view which is called ‘knowledge without belief’ (KWB). I apply this view to self-deception, and then demonstrate that KWB allows us to attribute knowledge to a self-deceiver’s intentional action without encountering the problems of Marcus’s view.

In the second paper, I explain different theories of the belief-acceptance distinction and argue that they fail to show what a self-deceiver’s belief is. The belief-acceptance distinction generates a view of self-deception which states that a self-deceiver holds an unpleasant justified belief that not-P, in spite of denying this due to accepting a pleasant unjustified proposition that P. This claim has two parts: a self-deceiver believes a proposition that not-P without accepting it, while accepting a proposition that P without believing it. Let’s call the first part ‘the belief side’ and the second part ‘the acceptance side’. After explaining the puzzle of self-deception and giving an overview of the literature on the belief-acceptance distinction, I object to this distinction by first arguing that the view
can only show how it is *epistemically rational* to believe a proposition without accepting it. Hence, it cannot account for a self-deceiver’s intellectual dishonesty. Then, I demonstrate that neither side of the argument works to solve the puzzle of self-deception, despite Erik Funkhouser’s (2019) claim that the belief-acceptance distinction is the best strategy to solve the puzzle at issue.

And in the final paper of my dissertation, I talk about epistemic blame. Different theories of epistemic blame are influenced by different accounts of moral blame. According to the emotion-based view of epistemic blame, inspired by the emotion-based views of moral blame, epistemic blame occurs when a blaming agent expresses emotional reactive attitudes such as indignation and resentment towards a blamed agent who has violated an epistemic norm culpably.

While I defend the idea that blame of any kind (practical, moral, epistemic) must take the form of emotional responses to a target of blame, I disagree that indignation and resentment are an appropriate emotional response to a target of epistemic blame. Hence, I revise the emotion-based theory of epistemic blame by replacing resentment and indignation with the emotion of disappointment or discontent. The reason for this claim is that a target of epistemic blame is different than a target of moral and practical blame, and a more minimal conception of blame suffices to account for epistemic blame. Thus, we may appeal to a variable theory of blame according to which blame may take different forms, and it is not necessarily a manifestation of the emotional reactive attitudes of indignation and resentment.
First Paper

How to Reconcile the Knowledge Thesis with Self-Deception

1) Introduction

The Knowledge Thesis (KT) is a view which says that it is impossible for an agent to act intentionally without knowing what he is doing. KT is advanced by Elisabeth Anscombe (2000), and Eric Marcus (2019) defends this view against self-deceptive action: a phenomenon in which an agent sincerely disavows knowledge of his intentional action all the while knowing what he is doing. KT’s opponents deem self-deception to be one of the most compelling putative counterexamples of KT, since a self-deceiver acts intentionally without knowing what he is doing. Both Marcus and the opponents agree that a self-deceiver’s act is intentional. What they disagree about is the component of knowledge. While the opponents don’t attribute knowledge to a self-deceiver, Marcus argues that a self-deceiver knows what he is doing, but masks his knowledge. In terms of Marcus’ view, a self-deceiver’s knowledge must be masked to justify distinguishing him from a bald-faced liar (i.e. an agent who is conscious of what he is doing intentionally despite denying it) when both agents similarly disavow knowledge of what they are doing.

Since Marcus is going to make a reconciliation between self-deception and practical knowledge, he only deals with self-deceptive action and not (what may be called) a self-deceptive fact: those cases of self-deception in which agents deceive themselves about facts related to their significant concerns. These facts, on Marcus’ judgment, are “external to the exercise of one’s agency” (e.g. deceiving yourself about whether your child is intelligent. Your spouse is faithful, etc.) (Marcus, 2019, p. 2).
Relating back to self-deceptive action, take this as an example of it: Gil, a self-deceived husband, keeps separating his wife, Maya, from their handyman, Johnny, by taking over landscaping which is Johnny’s job when he comes over to Gil’s house. Gil is doing landscaping to prevent the handyman from coming to his house and having an affair with his wife. Nevertheless, he denies that he is doing so due to that reason, and appeals to another reason to explain why he is doing the job.

And here is another example of self-deceptive action: an older brother finishes his younger brother’s sentence to humiliate him, but he states that he is only trying to help his brother. According to Marcus, in contrast with a self-deceptive fact, all cases of self-deceptive action involve a self-deceiver’s exercising his agency.

To define practical knowledge, Marcus refers to Anscombe’s definition. According to this view, a self-deceiver’s being able to just say what one is doing suffices to attribute practical knowledge to the self-deceiver, and there is no requirement for a self-deceiver to have that ability on the basis of observation or evidence.

In terms of Marcus’ view, a self-deceiver is a person who is able to say what he is doing when he is acting intentionally. Marcus also extends this view to those cases in which an agent has the ability in question, but the ability is masked due to different reasons. For instance, you might be able to ride a bike, but this ability is masked due to different reasons (e.g. being drunk, having twisted muscles, being hit by a rock in your way, being mentally preoccupied, etc.). Marcus uses this view to show how it is possible to attribute practical knowledge to a self-deceiver who has the ability to say what he is doing, although he cannot just say so, due to having the ability masked for one or a variety of reasons.
KT’s opponents believe that since a self-deceiver cannot just say what he is doing, then he doesn’t have practical knowledge of his intentional action. For example, Gil cannot (and so is not able to) say what he is doing when he is separating his wife from Johnny. As a result, he acts intentionally without knowing what he is doing.

Nonetheless, Marcus thinks that in the case of self-deception, the sense of ‘can’t’ doesn’t imply an absence of the ability to say so. An ability’s absence in general is only one explanation among different reasons why a person cannot perform an act. The fact that a self-deceiver can’t explain what he is doing comes from the ability’s being masked and not from lack of the ability. For instance, it is too painful for Gil to admit that he is worried about an affair. Hence, he cannot say that he is separating Maya from Johnny. But not being able to just say so doesn’t indicate Gil’s lack of the ability to know what he is doing. It only shows that Gil masks his ability due to being afraid of the truth.

To define the component of intentionality, Marcus argues that a self-deceiver’s act is intentional in a sense that Anscombe suggests: an act is intentional when a “why” question has an application here. For instance, in the case of the two brothers, the older brother finishes his younger brother’s sentence in fact in order to humiliate him, although non-lyingly disavows the correct interpretation of his behavior by saying that he aims to help his brother.

According to Marcus’ view, since the older brother’s masked knowledge is unconscious, then he is not a bald-faced liar when he sincerely denies the correct interpretation of his behavior. In addition, resorting to masked knowledge makes it possible for Marcus to argue that a self-deceiver is a person who has contradictory beliefs that P and not-P without having them in mind at once. Due to different reasons (e.g. distraction,
confusion, incomprehension, repression, etc.), one of the contradictory beliefs is occurrent or conscious, whereas another one is standing or unconscious.

Having laid out Marcus’ idea, I will argue that Marcus’ view is flawed by showing that it cannot truly capture the phenomenon of self-deception, and so fails to successfully defend KT.

2) Marcus’ Failure to Show Self-deception’s Tension

The first problem with Marcus’ view is that this view cannot show self-deception’s tension, despite showing that a self-deceiver is a person whose verbal assertion conflicts with his behavior. The tension here involves some hope that a proposition that \( P \) is the case, and simultaneously some doubt and/or suspicion that the proposition that not-\( P \) is the case too. The former proposition is normally associated with a pleasant but unjustified fact, whereas the latter one refers to an unpleasant but justified fact.

As is mentioned, Marcus thinks that a self-deceiver possesses masked knowledge, and this causes his knowledge to remain at an unconscious level. We also saw that holding unconscious knowledge makes it possible for an agent to have contradictory beliefs that \( P \) and not-\( P \) without having them in mind at once. For the sake of argument, I assume that masked knowledge is an example of unconscious knowledge. I am also aware that nothing is wrong with this kind of contradictory belief attribution to a normal agent in general. However, this attitude towards a self-deceiver’s belief is problematic, since it reduces the problem of self-deception to a simple case of holding two contradictory beliefs in which one contradictory belief appears after another one.
In other words, the problem for Marcus is that if the knowledge is standing and stored in a self-deceiver’s mind, then there is no tension for a self-deceiver. In order for an account to successfully explain self-deception’s tension, the account must show that a self-deceiver is conscious that a proposition that not-\(P\) is the case, while getting himself to believe a proposition that \(P\). Take Gil as an example again. Gil will suffer from no tension if he is not conscious that his verbal assertion (e.g. I take over landscaping, since I like it or I want to help Johnny) is a denial of what he is doing intentionally (i.e. taking over landscaping to separate his wife from another man). At most, Gil, in terms of this view, would be a person with a lack of enough self-conscious knowledge\(^1\) who doesn’t sufficiently reflect on his knowledge and behavior. Hence, he is unconscious of their potential contradiction, as opposed to a self-deceiver who has conscious awareness of a conflict between his verbal and nonverbal behavior.

To tackle my objection, Marcus would say that nothing will be wrong if we assume that the puzzle of self-deception lies in one’s non-lying denial of self-conscious knowledge:

What action, belief, and feeling all (arguably) have in common is this: there is some kind of constitutive connection between acting, believing and feeling, and knowing oneself to act, believe, and feel in a distinctively first personal and so non-evidential way. One might call this knowledge self-conscious knowledge. We could then say that self-deception- of this family anyway- amounts to a non-lying denial of self-conscious knowledge. If that is right, then perhaps the fundamental puzzle of self-deception is not, as it is arguably held, the fact that one believes against evidence, but that one is alienated from one’s self-conscious knowledge (Marcus, 2019, p. 28).

\(^1\) One might think that the notion of ‘self-knowledge’ suffices to show what we mean here, and there is no point to use the term ‘self-conscious knowledge’. While this point is true, I am going to use the latter term (which is used by Marcus) to be on the same page as Marcus.
In the above passage, Marcus tries to show that the puzzle of self-deception is more about one’s failure to know one’s action, belief, and feeling than an intellectual dishonesty a self-deceiver commits by believing against evidence. He argues for his claim by showing that the component of consciousness is not necessary for self-deception. I think this claim is wrong for the following reasons:

First, we tend *not* to hold a person with a lack of self-conscious knowledge *as* epistemically (or sometimes morally) *responsible as* an agent who is fully conscious of a contradiction between his belief and action, because the former is deemed to be *ignorant* of the contradiction (although we hold him to some extent responsible for his failure to reflect on his belief and action). But we hold a self-deceiver *as responsible as* an agent who is fully conscious of the contradiction, or at least we think that a self-deceiver is more comparable to this agent than an agent with a lack of self-conscious knowledge, since a self-deceiver *sees* evidence against an unjustified but pleasant fact, but biasedly ignores it. If consciousness were not of crucial importance in self-deception (as Marcus thinks), then a self-deceiver would be *as responsible as* an agent with a lack of self-conscious knowledge when he contradicts himself. But we saw that we hold the former more responsible than the latter.

Second, in many cases, an agent with a lack of self-conscious knowledge tries to modify his behavior or belief system after being conscious of evidence showing their contradiction, whereas a self-deceiver biasedly ignores the existence of unpleasant evidence in spite of being conscious of it or denies that there is such a contradiction between his behavior or belief system. If, as Marcus believes, the problem of self-deception is more about one’s being alienated from one’s self-conscious knowledge than believing
against evidence, then a self-deceiver, the same as the other agent, must modify his behavior or belief after being conscious of their contradiction. But this is not what a self-deceiver does.

Third, any act of deception (whether interpersonal deception or self-deception) involves both a deceiver (an active agent) and a person who is deceived (a passive agent). A self-deceiver is on a par with an interpersonal deceiver who intends to deceive another one. While a deceiver needs to be conscious of what he is doing, a deceived agent must remain unconscious, otherwise he would not be deceived. But eliminating consciousness from self-deception weakens the component of active agency which is half of the phenomenon of self-deception.

Fourth, in the case of self-deception consciousness exists despite a self-deceiver’s masking his knowledge. Masking is an intentional act that requires an agent’s conscious knowledge of what he is doing (i.e. the act of masking) and what is going to be masked as well. Masking one’s knowledge in this sense is similar to one’s hiding something from oneself or stealing out of one’s own pocket. One cannot hide something from oneself, unless one is conscious of the act of hiding and the hidden thing. In the next section, I will develop this point more in detail.

3) Masked Knowledge and the Problem of Consciousness

The second problem with Marcus’ view is about Marcus’ explanation of a self-deceiver’s masked knowledge as an instance of unconscious knowledge. To defend the knowledge thesis, Marcus argues that a self-deceiver’s intentional act is infused with knowledge. But in order to explain how a self-deceiver non-lyingly disavows knowledge of his intentional
action, Marcus tries to show that a self-deceiver’s masked knowledge is unconscious. Marcus’ view encounters a problem that I am going to call ‘the problem of consciousness’ (i.e. a self-deceiver’s masked knowledge cannot be unconscious, since the agent is aware of masking). Marcus is aware of a problem of this kind when he introduces a couple of potential objections to his view. Having introduced these objections, I aim to demonstrate that Marcus’ responses to them are not satisfactory.

According to the first objection, “if the agent must intentionally put her practical knowledge out of mind in order to perform a self-deceptive action, then the self-deception is itself intentional and known to the agent” (Ibid, p. 25). I think this objection is going to show that a self-deceiver’s knowledge remains conscious even after being masked. To make this point, this objection relies on the assumption that the components of intentionality and consciousness are tied together in self-deception: since a self-deceiver intentionally puts his knowledge out of mind, he is conscious of what he is doing, and so self-deception is known to the agent. To tackle the above objection, Marcus says that:

But since my thesis is precisely that one can know what one is doing without having it in mind, I can allow that he also masks his knowledge that he is masking. One can intentionally, for instance, evade contemplating an unpleasant fact (and so know that one is doing it) without bringing the fact that one is evading it to consciousness (p. 25).

In his response, Marcus suggests two solutions: he makes a distinction between intentional and conscious actions, and applies this distinction to self-deception. He also comes up with a view that I am going to call ‘second-order masked knowledge’ (i.e. a theory which says that an agent masks his knowledge that he is masking).
Now I am going to investigate Marcus’ distinction between intentionality and consciousness. Marcus’ point shows that his attitude towards intentionality is different than his opponents’ attitude. Marcus thinks that what makes an act intentional is derived more from one’s reasons for action than the component of consciousness. According to Marcus, when an agent masks his practical knowledge, he does that intentionally which means that he has a reason why he is doing so. Otherwise, masking would be done by accident, or it would be a byproduct of another action.

Nonetheless, to value the component of consciousness, Marcus doesn’t deny that at some point a self-deceiver (e.g. an older brother when he finishes his younger brother’s sentence) must have made the connection between what he does (e.g. finishing his brother’s sentence) and the real implication of that action (e.g. humiliating his brother), since a self-deceiver must be aware of the implication of his behavior. Otherwise, he would be comparable to a neglectful person who inattentively has made a mistake without knowing what his behavior indicates.

According to Marcus, the older brother has conscious knowledge of his behavior’s implication, but continuously suppresses the knowledge by saying that he is genuinely motivated to help his brother when he finishes his sentence. Marcus thinks that since the knowledge is suppressed, then the initial connection between a self-deceiver’s intentional act and consciousness doesn’t cause a problem for his idea that a self-deceiver’s masked knowledge is unconscious.

Nevertheless, I think Marcus’ view is problematic, since the mechanism of suppressing knowledge here is not clear enough. If a self-deceiver willfully suppresses his knowledge, then he is still aware of what is suppressed, and the knowledge cannot go to an
unconscious level. Even an appeal to second-order masked knowledge cannot solve the problem, since even if a self-deceiver successfully masks his knowledge that he is masking, he is still *conscious* of the so-called second-order masked knowledge, and as a result, the knowledge in question cannot be standing or unconscious.

It seems that to resolve those problems threatening the first-order masked knowledge, Marcus comes up with the second-order masked knowledge. But all of the problems remain at the second level too. In other words, to tackle the problem of consciousness, Marcus says that his view allows that a self-deceiver masks his knowledge that he is masking. By this logic, Marcus must also allow for the existence of third-order masked-knowledge, fourth-order masked-knowledge, and so forth, since consciousness still exists at any of these levels, and in order to mask it, we need to have a higher-order masked knowledge. But this will never end, and will be problematic.

Relating back to the relation of intentionality with conscious and unconscious knowledge, it’s worth noting that in the case of standing or unconscious knowledge, one *doesn’t decide not to have* the knowledge in mind. That is, it’s almost beyond one’s control when one’s occurrent knowledge goes to an unconscious level. These cases occur when an agent doesn’t need to reflect on his knowledge in a continuous conscious way. For instance, when an agent repeats the same intentional process for a long while (e.g. driving a car, brushing teeth, etc.), doing the process becomes his mental habit, and thus the agent is not required to be totally conscious of what he is doing all the time.

But masking is different than the above-mentioned process. In order for masking to be done, there is no need to perform a repetitive behavioral pattern which occurs frequently. Hence, masking doesn’t normally go to a subconscious level in which one’s continuous
conscious reflection may be legitimately set aside. Therefore, I think Marcus is wrong in his claim that in the case of masking intentionality and consciousness go in different directions.

However, I don’t deny that, comparable with any other act of hiding, an agent might forget what is masked as time goes by. For instance, in the case of the two brothers, after a long while the older brother may forget his old intention (i.e. humiliating his brother). But if forgetfulness occurs, then another phenomenon will be born here: the agent at most will be a person with a lack of self-conscious knowledge again without feeling any tension. As is mentioned earlier, it would be difficult to call this phenomenon self-deception.

I think Marcus’ opponents are right in their claim that intentionality and consciousness are inextricably intertwined when it comes to masking. We also saw that there are other cases in which intentionality and consciousness go in different directions (e.g. driving a car). Since these cases are fundamentally different than masking, the distinction between intentionality and consciousness cannot play any role in a definition of masked knowledge. If Marcus is still going to show that in self-deception intentionality and consciousness go in different directions, he must appeal to a different account than masked knowledge.

What is problematic for Marcus’ view is that on the one hand, Marcus has to acknowledge that a self-deceiver (at least at some point) has conscious knowledge of his act, and on the other hand, in order to distinguish a self-deceiver from a bald-faced liar, he thinks that this knowledge must remain unconscious. But, as I said, a self-deceiver’s knowledge of his act cannot and should not be unconscious. I believe that there is another
way to successfully distinguish a self-deceiver from a bald-faced liar without appealing to unconscious knowledge\(^2\). I will talk about this in the next section after presenting another problem for Marcus’ view.

4) Self-deception and Different Levels of Knowledge

Marcus talks about another potential problem that his opponents may come up with. This problem is somewhat similar to the problem of consciousness, and it assumes a variation of the principle discussed in the previous section: “\(y\)-ing intentionally in order to \(x\) entails knowledge that \(y\)-ing is a means of \(x\)-ing” (p. 25). In other words, one sees that \(x\)-ing is an upshot of \(y\)-ing. And if one sees this causal relationship, then one’s knowledge that one is intentionally \(x\)-ing is conscious and present in one’s mind.

Marcus states that this objection fails even if for the sake of argument we assume that one sees that \(y\)-ing would bring about \(x\)-ing. He offers this analogy to make his point: suppose that you play music loudly late at night and you see that by doing so, you annoy your neighbor. We can rightly assume that you are not annoying your neighbor intentionally, since playing the loud music is the only way by which you can shake the bomb off the bathroom sink and into the tub full of water. In this situation, you don’t intentionally annoy your neighbor, even if you are conscious that annoying your neighbor is a consequence of what you are intentionally doing. Cases like this, in Anscombe’s

\(^2\) In addition, it is not determined how a self-deceiver will move back and forth between a conscious and unconscious level and how these two levels will interact with each other if a self-deceiver’s mind is partitioned off between the two distinct realms. Marcus’ view reminds us of Davis Pears (1984) and Donald Davidson (1985) when they appeal to the Freudian idea that the best way to solve the puzzle of self-deception is splitting the person. But, as José Eduardo Porcher (2014) says, these philosophers must answer how a self-deceiving agent will be capable of deceiving when there is a block between the two parts of his mind (Porcher, 2014, pp. 296-297)
language, are called voluntary, but not intentional (Anscombe, 2000, pp. 89-90). Marcus thinks that Gil’s case (when he separates his wife from another man) is analogous to the above case:

So even if he is conscious of the fact that he is keeping them separated, his knowledge that it is *for the sake of that end* that he performs the means-action can nonetheless be masked, and thus so can his knowledge that he is intentionally performing the end action. He sees that y-ing will promote x-ing, but he does not see that he is y-ing in order to x (Marcus, 2019, p. 26).

I believe that Marcus’ reply to this objection is not satisfactory either. It’s not determined why a self-deceiver is conscious of the former fact (i.e. separating his wife from another man), but he is unconscious when it comes to the two latter ones (i.e. the fact that it is *for the sake of a specific end* that he performs a means-action and the fact that he is intentionally performing the end action).

I am aware that there is a difference between Gil’s knowing the former and the two latter facts in terms of dealing with an unpleasant truth: whereas the two latter facts obviously deal with an unpleasant truth that the agent tries to avoid thinking of, the former merely refers to a fact that can be neutral in terms of containing a pleasant or an unpleasant truth. That is, Gil’s separating his wife from another man (or in this example: the older brother’s finishing his younger brother’s sentences) is a mere act comparable with many actions we do in our daily life. These actions are a chain of events occurring in the world without necessarily conveying any positive or negative value judgment. It’s one’s further reasons for doing an act that attach value to it and make the act pleasant or unpleasant. For
instance, feeling jealous is an actual reason that motivates Gil to perform the act of separating his wife from Johnny. This act can be pleasant if Gil sees it as a way of being more secure, and unpleasant if it reminds him of a potential threat. The act by nature doesn’t convey any essential positive or negative value regardless of the context in which it is considered.

What is always unpleasant is Gil’s reason (i.e. feeling jealous) for separating his wife from another man. Since this reason is not pleasant, Gil tries to ignore his knowledge that it is for the sake of that reason he separates his wife from another man. And for the same reason, Gil also aims to disregard his knowledge that he intentionally performs the act in question.

But the point is that the difference between the former and the two latter facts is merely about being pleasant and unpleasant, and thus the difference is not so fundamental that renders one’s knowledge of the former conscious and the two latter ones unconscious.

Marcus would answer that one’s knowledge of the former is fundamentally different than knowledge of the latter ones, since when it comes to the former, all one needs to obtain is a surface level of knowledge which is merely about describing what one is involved in. By contrast, the two latter cases deal with a deeper level of knowledge. This knowledge is more difficult to obtain, since it is about discovering and interpreting an act’s hidden reasons or ends. Hence, it requires an agent to spend more time and effort to gain the knowledge.

For the sake of argument, I assume that Marcus’ above claim is right. But I believe that the fundamental difference between the two types of knowledge is an issue when an
agent *is not aware* of potential motivating reasons influencing his action. Consequently, the agent must investigate more to realize his hidden reasons. This doesn’t happen in self-deception, since a self-deceiver has *conscious awareness* of his reasons for action, in spite of denying the reasons due to their unpleasantness.

I am moving to a slightly different point. In the above-cited passage, Marcus states that a self-deceiver sees that y-ing will promote x-ing, but he does not see that he is y-ing in order to x. This statement can be interpreted in two ways: if a self-deceiver doesn’t really see that he is y-ing in order to x, then a self-deceiver *doesn’t intend* to x. This claim, if it is true, would be against the fundamental assumption that self-deception is intentional. That is, based on this interpretation, we must conclude that, for instance, the older brother doesn’t intend to humiliate his brother. But I don’t think that Marcus is willing to accept this interpretation which rules out his own assumption about the intentionality of self-deception.

Hence, the only way for Marcus is accepting the second interpretation: a self-deceiver has the intention of x-ing though he is unaware of his intention. To tackle the problem of the first interpretation, Marcus would have to argue again that the intentionality at issue does not need to be discernible for a self-deceiver. He would probably suggest again that at some point a self-deceiver has seen that he is y-ing in order to x, but he masks his knowledge of what he is doing. As a result, it is legitimate to attribute the intention of x-ing to him. But Marcus’ potential suggestion here would encounter again the problem of consciousness: a self-deceiver’s masked knowledge cannot remain unconscious. And if the knowledge is conscious, then Marcus needs to answer how he distinguishes a self-deceiver from a bald-faced liar who is also conscious of the truth.
Marcus doesn’t have any answer for this question. I think the difference between a self-deceiver and a bald-faced liar is that while both agents have conscious knowledge of a fact they deny, they show a different doxastic response to the denied fact. For example, in contrast with Gil, a bald-faced liar does believe that his wife has had an affair, despite denying the fact in front of others. But Gil biasedly ignores evidence indicating his wife’s infidelity so as not to believe the unpleasant fact, although he is conscious of it. And this is why we think that Gil commits intellectual dishonesty, since he refuses to believe in spite of having enough evidence.

In the next section, I will introduce a new view to support my claim, and then show how the view tackles those problems that Marcus’ view faces.

5) Knowing without Believing

I believe that the theory of ‘knowledge without belief’ (KWB) approves of the claim that a self-deceiver has conscious knowledge that not-\(P\) without believing it. KWB has been originally advanced by Radford (1966) and has been resuscitated by Myers-Schultz and Schwitzgebel. The advocates of this theory believe that it is intuitively correct to claim that a subject knows some proposition that \(P\) without any belief (or at least determinate belief) that \(P\). This case is a paradigmatic example of KWB: Kate has studied well for her history exam, and everything goes well during the exam until her teacher rushes students to wrap up. The final question is this: “when did the Queen Elizabeth die”? This is a question that, before the exam, Kate had studied many times, and recited to her friends too. During the exam, she also sees the question, and is confident that she knows the answer. However, having been rushed by her teacher in the last minute, she starts getting worried and that causes her not to recall the answer anymore. Disappointedly she writes “1603” as the
answer which happens to be the right answer. According to the advocates of knowledge without belief, two claims follow from the example:

1) Kate *knows* that “Queen Elizabeth died in 1603”, since she has answered many questions, including the question at issue, correctly as the result of studying them.

2) Kate does not believe that “Queen Elizabeth died in 1603”, since she thinks that her correct answer was a mere guess, and she doesn’t take it to be true\(^3\).

This is an example originally posed by Radford, and it is modified by Myers-Schultz and Schwitzgebel. Myers-Schultz and Schwitzgebel refer to the result of some research in which respondents (i.e. normal speakers) recognize different scenarios similar to the above example as a case of knowledge without belief\(^4\). This sort of attitude towards knowledge reminds us of Williamson’s generality claim:

knowledge need not be a belief state that satisfies certain epistemic constraints.

Rather knowledge is a determinable of which other mental states are determinates. Perceptual states, standing belief states, judgments, realizations, recollections, ability states, introspective states, and so on are all determinates

\(^3\) A defender of the traditional view on knowledge (which considers belief as a component of knowledge) may distinguish between dispositional and occurrent belief and also dispositional and occurrent knowledge here. For instance, Kate knows and so believes in a dispositional sense, although she lacks the occurrent knowledge or belief. This explanation may be true for cases in which one is, for instance, in an unconscious state or doesn’t have access to stored knowledge or belief. However, there are cases in which one has conscious awareness that something is the case, while talking or acting as if one doesn’t know or believe it. These cases cannot be explained by the traditional view. Self-deception belongs to this group, and I think KWB is the best explanation of it.

\(^4\) One of these scenarios, a self-deceived husband, is exactly about self-deception and reminds us of Gil’s case. In this example, the husband sees strong evidence indicating that his wife has cheated on him, however, he keeps denying that. The result has shown that many normal speakers have voted that the agent knows that he is cheated on without necessarily believing it. Although this finding is very influential for solving the puzzle of self-deception, Myers-Schultz and Schwitzgebel haven’t developed and classified it as an independent view. This is a task that I am going to do.
of knowledge, as long as they satisfy certain epistemic constraints. Some of these, for example, seeings, are primitive knowledge states. Others are standard knowledge states (Williamson, 2002, cited in Bengson & Moffett, 2011, P. 152)

This attitude (i.e. knowledge without belief) has different influential implications for self-deception. To show how KWB is applicable to self-deception, I need to show first what “knowledge” and “belief” indicate in terms of KWB. In the example of Kate, we attribute knowledge to her, since she has acquired true information in a justified way, despite not recalling it. Nonetheless, we don’t think that she believes it, because she is not committed to what follows from taking a proposition to be true. Since she takes her correct answer as a mere guess, her internal state is different than someone who will never be able to give up her belief that, for instance, “Earth orbits the sun”, even if she is offered millions of dollars to believe the other way.

Kate’s case is different than a self-deceiver’s case somewhat, in terms of knowledge possession, but this difference, that I am going to explain now, doesn’t prevent us from regarding these two cases as different examples of KWB. In Kate’s case, the examinee doesn’t recall the answer, which means that she is not conscious of it, whereas in self-deception (e.g. Gil’s case) the agent is aware of the existence of unpleasant evidence through a justified way and what information the evidence indicates. But, unlike Kate, the self-deceiver ignores or manipulates evidence in order not to acquire new belief⁵. However,

⁵ On might claim that this line of reasoning wrongly adopts doxastic voluntarism (i.e. a view which says that agents voluntarily form or refuse to form their belief), because the belief is already formed in spite of the self-deceiver manipulating evidence. This is a potential objection that I will answer in the next pages.
knowledge is attributable to the self-deceiver in terms of KWB, since the self-deceiver stands in an epistemic relation (i.e. a perceptual and a memory state in this example) with the relevant evidence that justifies attributing knowledge to him.

This kind of knowledge attribution has been discussed by other epistemologists as well. Bengson and Moffett (2011) have the same view towards knowledge when they revise Ernest Sosa’s (2007) distinction between animal knowledge and reflective knowledge. According to Sosa, reflective knowledge requires an agent to hold reliable second-order belief, whereas in order to have animal knowledge, an agent’s belief must be apt (i.e. an accurate belief obtained by an agent’s exercising an epistemic competence). Bengson & Moffett argue that one can have animal knowledge without necessarily the apt belief requirement: “instead of saying that one has animal knowledge just when \( s \) has an apt belief, one could say that \( s \) has animal knowledge that \( p \) when \( s \) is in some state with the content of ‘\( p \)’ that is apt (e.g. a perceptual state, a memory state, an introspective state, a belief state, and so on)”. (fon & Moffett, Ibid).

Since KWB attributes knowledge to a self-deceiver’s intentional act, it shows how self-deception and KT are compatible. Another implication of KWB for the puzzle of self-deception is that the theory also allows us to successfully distinguish a self-deceiver from a bald-faced liar: a self-deceiver knows a certain (unpleasant) fact without believing it as opposed to a bald-faced liar who both knows and believes the fact.

KWB attributes knowing that not-\( P \) to a self-deceiver (without requiring him to believe it) and simultaneously allows him to hold a belief that \( P \). Since a self-deceiver doesn’t necessarily believe not-\( P \), then this account doesn’t face the problem of contradiction (namely, the problem of holding two contradictory beliefs simultaneously).
We saw that on Marcus’ view knowledge entails belief. As a result, Marcus needs to tackle the problem of contradiction when he attributes contradictory beliefs that $P$ and not-$P$ to a self-deceiver. So, he appeals to masked knowledge to show how one belief is masked or unconscious as opposed to another contrary belief which is conscious. But, as is mentioned earlier, a self-deceiver’s knowledge cannot and should not be unconscious.

However, an opponent of my view might say that this transition from a doxastic level of a belief that not-$P$ to an epistemological level of knowledge without belief cannot capture the tension imbedded in self-deception, since in order for the tension to exist, possessing two contradictory beliefs is required.

My answer to this objection is that the tension at issue doesn’t necessarily need to be caused by holding two contradictory beliefs. The tension is noticeable when an agent is conscious of the existence of unpleasant evidence which makes him suspicious that something unpleasant is the case (although no outright belief has been formed in this level yet), but he is afraid of investigating the case more, despite having an epistemic obligation to do that. And this is what we need to show the tension of self-deception.

The way KWB explains the tension at issue can also show why a self-deceiver commits intellectual dishonesty, because one might ask if knowing that $P$ doesn’t necessarily entail believing that $P$, then it seems that a self-deceiver hasn’t necessarily committed any intellectual dishonesty (like Kate in the example of an unconfident examinee), while we know that self-deceivers have violated an epistemic obligation. In other words, we don’t deem Kate to be irrational, despite knowing that “Queen Elizabeth died in 1603” without believing it. But the same doesn’t go for a self-deceiver. How does KWB account for a self-deceiver’s intellectual dishonesty? Or what is unique about a self-
deceiver, compared to Kate, that makes self-deception an epistemically irrational phenomenon?

This question can be answered by looking at the self-deceiver’s manipulating evidence to relieve his tension. A self-deceiver knows something that epistemically requires him to go further and believe not-\(P\). Nonetheless, the agent tries to convince himself that \(P\) by manipulating evidence or relying on irrelevant evidence. And this is what makes self-deception epistemically irrational.

Now I am going to pose another potential objection against my view. One might object that KWB allows for the illegitimate view of doxastic voluntarism (namely, a view which says that agents form or refuse to form their belief voluntarily). According to this objection, when a self-deceiver is subject to unpleasant evidence, his belief that not-\(P\) is already formed, despite the agent’s denial of it. But KWB allows a self-deceiver not to form a belief that not-\(P\) voluntarily, whereas we know that belief formation is not a matter of choice. If this objection is correct (and the belief that not-\(P\) is formed involuntarily), then we have enough reason to attribute the possession of the belief that not-\(P\) to a self-deceiver, since the evidence of this belief is revealed to the self-deceiver, and as a result, the self-deceiver cannot help holding the belief.

This objection doesn’t work. It is not that a self-deceiver voluntarily refuses to form a belief that not-\(P\). In fact, a self-deceiver is forced by his emotions not to form a belief that not-\(P\). Lazar also mentions this point when he argues for the role of emotions in self-deception:

Emotions don’t affect one’s view of the world through deliberation: they do so immediately and in a way which, to a high degree, is not subject to our control.
Fear, hope and their kin (and in concern with other elements affect cognition without the existence of a plan that is designed for this purpose (Lazar, 1999, p. 282).

In addition, another problem with this objection is that it advances a normative claim: it’s an agent’s epistemic responsibility to form belief in the presence of evidence. I think this epistemic responsibility deals with a normative principle that any rational agent should follow, whereas what I am going to show here merely describes how an irrational agent forms or refuses to form belief. There is a difference between what an agent should do and what an agent does, and my solution to the puzzle of self-deception deals with a descriptive and not a normative claim. Additionally, my view doesn’t claim that it is epistemically rational for a self-deceiver not to form a belief that not-\(P\) due to the strong force of her desires and emotions. All I show is why a self-deceiver, an irrational agent, is motivated not to form a belief that not-\(P\).

**Conclusion:**

In order to show how the knowledge thesis and self-deception are compatible, Marcus argues that a self-deceiver possesses masked knowledge. That is, a self-deceiver unconsciously knows a proposition that not-\(P\) while non-lyingly asserting a proposition that \(P\). But I showed that this view is problematic for different reasons. Marcus’ view wrongly treats a self-deceiver the same as a person with a lack of self-reflection. In addition, I showed that masked knowledge cannot be unconscious, since one cannot intentionally make one’s conscious knowledge unconscious. I also demonstrated that Marcus’ view fails to account for a self-deceiver’s tension, since it cannot show that a self-deceiver has conscious awareness of a proposition that not-\(P\) when he sincerely asserts a
proposition that $P$. I argued that due to this flawed account of self-deception, Marcus cannot successfully demonstrate how self-deception and the knowledge thesis are compatible.

After showing why Marcus’ view doesn’t work, I used the theory of ‘knowledge without belief’ to show the compatibility of self-deception with the Knowledge Thesis. I argued that the theory at issue may successfully attribute knowledge to a self-deceiver without requiring him to necessarily believe that not-$P$. Since a self-deceiver doesn’t hold a belief that not-$P$, then we don’t need to resort to masked knowledge or unconscious belief to reconcile holding a belief that not-$P$ with a belief that $P$. However, this view allows us to explain a self-deceiver’s tension by showing that a self-deceiver is conscious of an unpleasant fact (which makes him suspicious that the unpleasant fact is the case) while trying not to believe it or getting himself to believe the other way.
1) Self-Deception

According to the \textit{intentional interpersonal model} of self-deception (known as the \textit{traditional account}), self-deception is a phenomenon in which an agent gets herself to believe that $P$, while she knows or believes that not-$P$. In terms of this model, the phenomenon has two features: a) It is intentional, namely, agents intentionally get themselves to actually believe a proposition, all the while knowing or believing that it is false (Call this ‘the intention requirement’). This feature makes a distinction between self-deception and a mere error, where the latter is caused due to some accidental or unintentional mistake in the process of forming or maintaining belief. b) Self-deceivers hold two contradictory beliefs simultaneously and consciously (Call this ‘the belief requirement’).

This view of self-deception results in a puzzle named \textit{‘the static paradox’} which is about one’s state of mind: how is it possible for a person to hold two contradictory beliefs consciously at the same time (perhaps an impossible state of mind)?

One of the prevalent views to solve the puzzle in question is this: it is only not-$P$ (an unpleasant justified proposition) which represents a self-deceiver’s actual belief, and a proposition that $P$ (a pleasant unjustified proposition) is merely an agent’s sincere assertion. Robert Audi (1988) takes this side by arguing that a self-deceiver’s non-verbal behavior indicates the existence of an \textit{unconscious} belief that not-$P$ (namely, a type of
belief one has without being conscious of it. This sort of belief requires more effort to be discovered).

Johnathan Cohen (1992) also aligns himself with this view (i.e. deeming not-$P$ to be a self-deceiver’s actual belief). Cohen uses the belief-acceptance distinction to argue for his claim. To support his view, Cohen also appeals to the existence of unconscious belief to show how it is possible to attribute a certain belief to a self-deceiver in spite of her (verbal or non-verbal) denial. On this view, a self-deceiver doesn’t accept a belief that not-$P$, despite actually holding it. Instead, what a self-deceiver accepts, in terms of this view, is a pleasant unjustified proposition that $P$ without really believing it. Michael Bratman (1992) also offers an account of belief-acceptance distinction which seems to be more comprehensive than Cohen’s view, although Bratman doesn’t directly apply his theory to self-deception. Among different suggestions to solve the puzzle of self-deception, Erik Funkhouser (2019) argues that the belief-acceptance distinction is the most plausible view.

In what follows, in addition to presenting Cohen’s view, I am going to lay out Bratman’s account too as a supplementary view to give a comprehensive explanation of the belief-acceptation distinction. Then, I will argue why this distinction cannot solve the puzzle of self-deception.

2) Different Views of the Belief-Acceptance Distinction

In this section, I am going to combine Cohen’s and Bratman’s view of the belief-acceptance distinction and show in detail how this theory is applicable to self-deception. According to Cohen, in order to believe that $P$ one must be disposed to feel that it is true that $P$ and false that not-$P$ without being required to act, reason or speak accordingly. Bratman also counts
four conditions necessary for reasonable belief: first, belief is context-independent (i.e. belief doesn’t vary from context to context). Second, belief is formed based on evidence and is concerned about the truth of what is believed. Third, belief is also formed involuntarily. Fourth, belief is subject to an ideal of integration. That is, an agent’s belief should be consistent and coherent.

And, in terms of Cohen’s view, to accept a proposition that $P$ for use in inference, one must treat it as given that $P$. In other words, you accept that $P$ when you adopt a policy of deeming, positing or postulating that $P$. Moreover, by accepting a proposition, you include that proposition in your premises about what to do or think in a certain context without feeling whether the proposition is true or not.

Bratman supplements Cohen’s account of acceptance by adding more conditions to it: what we accept varies in different contexts for different practical reasons. Bratman states that practical pressures can make it reasonable to accept a proposition in a certain context and not in another one. Acceptance is sensitive to different contexts. It’s not that an agent has a context-independent attitude that keeps changing in different contexts. The point is that an agent can reasonably accept something in one context and reject it in a different context. For instance, when something doesn’t matter to you (or you only aim to simplify your daily plan), you just take it for granted that it is not supposed to rain on a normal day in July. But when you are supposed to place a bet on it, you don’t easily accept this. This example shows how different practical concerns urge us not to accept what we believe or the other way. In sum, according to Bratman, acceptance is related to context. It is under voluntary control. It is also guided by practical considerations, and it is not limited by holistic considerations of rationality.
Now let’s see how Cohen applies the belief-acceptance distinction to self-deception. Cohen doesn’t deny that there is a natural tendency for belief and acceptance to run together. However, there are cases in which these two states run in different directions, and this is the key to understanding the conceptual nature of self-deception. In terms of this view, in general, an unpleasant but justified belief that not-\( P \) is a belief that a self-deceiver actually holds, but represses and doesn’t accept it, while accepting a pleasant but unjustified proposition that \( P \) without believing it. For instance, a follower of a particular political party sees evidence against her favorite candidate’s honesty leading her to hold the belief that “the candidate is a fraud”. Nonetheless, she represses the belief, and doesn’t accept it. Instead, she accepts a proposition with a totally different content (e.g. “my favorite candidate is honest”). According to this view, a self-deceived person, merely accepts the latter proposition without believing it. The reason is that since a self-deceiver holds the unpleasant belief (i.e. the former proposition), she cannot hold its contrary pleasant belief (i.e. the latter proposition) simultaneously, since believing two contradictory propositions at the same time leads to an impossible state of mind. So, a self-deceiver holds only an unpleasant belief without accepting it, and merely accepts a pleasant proposition without believing it.

As I mentioned, unlike Cohen, Bratman doesn’t directly apply his theory to self-deception. But there is not going to be any remarkable difference between the outcome of applying his theory to self-deception and Cohen’s theory. Bratman’s view endorses Cohen’s idea when Bratman states that we don’t need to always rationalize our actions by finding a belief behind them, as the cognitive attitude directing our practical reasoning goes beyond our belief. For instance, you have a plan to do something. In your deliberation to
make your plan, there can be many propositions that you simply accept in the cognitive background of your deliberation without necessarily believing them. Or sometimes you believe (have a high degree of confidence in) a proposition but you reasonably decide not to accept it in an appropriate practical context. Cohen’s and Bratman’s view demonstrate how it is possible to believe a proposition without accepting it or simply accept a proposition without necessarily believing it.

In the next sections, I will show why the belief-acceptance distinction cannot successfully solve the problem of self-deception, despite Erik Funkhouser’s (2019) claim that the distinction between belief and acceptance is a good strategy to answer the puzzle.  

3) The Belief-Acceptance Distinction and its Failure to Show the Irrationality of Self-Deception

To begin with, I am going to mention that belief-acceptance is a distinction to show how it is *epistemically rational* for an agent to accept what she doesn’t believe or not to accept what she believes. However, to argue against their opponents (i.e. those who attribute only a belief that \( P \) to a self-deceiver), the proponents of the belief-acceptance distinction to solve the puzzle of self-deception use their theory to account for the *irrational* phenomenon of self-deception. And this creates a serious problem for this camp: a self-deceiver commits *intellectual dishonesty* by believing something but accepting something else as this view states, whereas the followers of this view cannot account for the irrationality of self-deception, since in order to analyze a self-deceiver’s belief, the proponents of the belief-acceptance distinction use the very account which they use to show how belief and acceptance can rationally go in different directions.
Nevertheless, there are two ways for the proponents in question to tackle the problem: either showing that self-deception is not an irrational phenomenon or accepting the irrationality of the phenomenon and trying to find a way to account for this irrationality. The former solution sounds too radical, since it will treat a self-deceiver (who is known to be biased, and so epistemically irrational) the same as an epistemically rational agent. And as Cohen states, any theory that ignores the irrationality at issue is doomed to failure.

Since no one gives credence to the first suggestion, then the only way for the advocates of the belief-acceptance distinction is finding a way to explain the irrationality of self-deception. Cohen tries to do this by showing how a self-deceiver treats evidence. In what follows, after laying out Cohen’s suggestion, I will explain why his view doesn’t work.

As Cohen states, it is not necessarily irrational to honestly accept a certain belief and drive a contrary belief out of your consciousness. If one is not successful in totally discarding the involuntary belief, then one can make it deactivated by accepting the opposite, and this process is not necessarily irrational. Nevertheless, Cohen is aware that his account, or any theory arguing for belief-acceptance distinction, must be able to show the intellectual dishonesty embedded in self-deception. According to Cohen, what makes self-deception irrational is misinterpreting or manipulating evidence:

The other distinctive feature of self-deceit that needs to be mentioned is the way in which the self-deceiver handles the evidence. And it is there that intellectual dishonesty or irrationality enters into the situation. No analysis that ignores this […] can be satisfactory. (Cohen, 1992, p. 144).
According to this view, a self-deceiver manipulates evidence either by failing to accept relevant premises for an undesired belief that not-$P$ (though the agent believes them to be both relevant and true) or by accepting premises for a desired belief that $P$ (though the agent doesn’t believe these premises).

I believe that even Cohen’s suggestion cannot show how a self-deceiver differs from a rational agent, since Cohen refers to a self-deceiver’s treatment of evidence (i.e. dismissing relevant evidence and/or considering irrelevant evidence), and this is what a rational agent does when she is going to withhold her actual belief for a while to accept its contrary proposition without actually believing it. In what follows, I will explain what distinguishes these two agents from each other is their different state of mind or attitude towards a particular proposition that the advocates of the belief-acceptance fail to account for. That is, I aim to show that while a rational agent holds or withholds a belief when it comes to a certain fact, a self-deceiver stands in a contrary doxastic relationship with that fact. I demonstrate that the belief-acceptance distinction fails to prove its claim (i.e. a claim according to which a self-deceiver believes an unpleasant justified proposition that not-$P$ without accepting it, while accepting a pleasant unjustified proposition that $P$ without believing it).

I think if we consider a self-deceiver and a rational agent *regardless of their intention*, then we will see that what a self-deceiver does (i.e. dismissing relevant evidence and/or considering irrelevant evidence) is what a rational agent does when she suspends her actual belief for a while to accept something different. For instance, a scientist, for the sake of argument, suspends her actual belief by dismissing its relevant evidence and accepting irrelevant evidence. As an external observer, if you are not aware of a scientist’s
intention, you will find this behavior comparable to what an irrational agent, a self-deceiver, does for her own purposes. Nonetheless, we know that there is a fundamental difference between the scientist, a rational agent, and a self-deceiver, an intellectually dishonest person, despite their similar behavior. But Cohen’s suggestion fails to show the difference.

This is why I state that referring to a self-deceiver’s treatment of evidence is not a good strategy to show the irrationality of self-deception, since the difference between a rational agent and a self-deceiver derives from the agents’ different state of mind: while a rational agent acknowledges to herself that her actual belief (i.e. what she takes to be true) is different than what she is going to voluntarily accept for a while (without really believing it), it’s very unlikely for a self-deceiver (due to seeming to be too painful) to admit to herself that “I only accept a proposition that $P$ without believing it, since my actual belief is a proposition that not-$P$”.

To show better how the two agents’ internal state differs, in what follows, I will explain what it takes for someone to believe and to accept something. Then, by means of this explanation I will demonstrate that Cohen’s and Bratman’s views of the belief-acceptance distinction don’t have any explanatory power to account for the irrationality of self-deception or they cannot successfully argue for any part of their claim: neither the belief part (namely, attributing a belief that not-$P$ to a self-deceiver without accepting it) nor the acceptance part (namely, attributing acceptance of a proposition that $P$ to a self-deceiver without believing it). I will argue for my claim by developing the belief part first.

4) The Belief-Acceptance Distinction and its Failure to Attribute a Belief that not-$P$ to a Self-deceiver
To show why a self-deceiver cannot be a believer of a proposition that not-$P$, I am going to explain first what it takes for someone to believe something. *Taking a proposition to be true* is the most proper condition of belief attribution. This condition is epistemically less demanding and less controversial than other conditions (e.g. those conditions that require an agent to reason and/ or act in terms of a belief in order to attribute a belief to an agent), since it doesn’t require an agent to do anything special to prove that she is a belief-holder. As a result, the condition is not so restrictive that it rules out those propositions that one takes to be true, while failing to act or reason in terms of them due to different reasons (e.g. phobia, weakness of will, prudential or practical concerns, etc.). On the other hand, it is not so broad that allows for one’s wishful thinking or desire to be taken as one’s belief, since it requires the norm of truth.

It’s worth noting that failing to act or reason in terms of a belief occurs either *involuntarily* (for instance, phobia forces people not to act in terms of their belief in spite of their own volition) or *voluntarily*: agents can intentionally and voluntarily avoid using their belief in their reasoning because of special concerns (e.g. a scientist, for the sake of argument, decides not to use her actual belief in her reasoning). In both cases, agents acknowledge that they take a particular belief to be true, and they are also aware that they would have reasoned or acted in terms of their belief in a *normal circumstance* (namely, a circumstance without any specific reason which might encourage an agent not to act or reason in terms of her belief for a while).

I stated that taking a proposition to be true is the best criterion for belief attribution. In addition to that, to hold a belief diachronically in a rational way, an agent is also
epistemically required to accept6 her belief (namely, postulating, acting, and/ or reasoning in terms of a belief) as well in a normal circumstance. That is, belief and acceptance must go hand in hand in a usual circumstance, and a rational agent usually accepts what she believes in a normal situation.

Since a rational agent is the one who genuinely takes a proposition to be true, this justifies us claiming that she believes the proposition. Even if for the sake of argument, the rational agent deactivates (namely, refraining from postulating, reasoning or acting in terms of a belief, despite taking it to be true) her actual belief for a while, we are still justified attributing the belief to her, since we know that she has deactivated her actual belief intentionally and temporarily, while still taking it to be true. But a self-deceiver doesn’t meet even this general condition. That is, she doesn’t even take an unpleasant proposition that not-\(P\) to be true under any circumstance. Hence, there is no way at all to conclude that a self-deceiver believes the proposition at issue7.

And when there exists no belief at all to begin with here, then it’s pointless to talk about accepting and then deactivating it. Deactivating a belief is what an agent must do in order to accept a proposition contrary to her belief (while still taking her belief to be true8), since no one can accept two contrary propositions simultaneously. And this is what a self-deceiver should do if, as the belief-acceptance distinction claims, she holds a belief that

---

6 In the final section, I will show that acceptance comes in two forms: voluntary and involuntary.

7 There might be an objection to this criterion of belief attribution that I will explain and answer in the next pages.

8 One might argue that deactivation is required only in the case of occurrent belief, but it doesn’t seem to be necessary when it comes to standing beliefs of which one is not aware. Although this point sounds to be true, it doesn’t cause a problem for the current debate on self-deception, since in self-deception we deal with conscious beliefs. The reason is that standing beliefs cannot demonstrate tension a self-deceiver suffers from.
not-\(P\), but accepts its contrary proposition that \(P\) without believing it. But the above explanation shows that the belief-acceptance distinction cannot even attribute to a self-deceiver holding a belief that not-\(P\).

It is apparent that in a normal condition, mere acceptance of \(P\) doesn’t necessarily require undergoing this process: \textit{believing not-\(P\) first, accepting and then deactivating it.} People might accept a proposition in their daily life without necessarily being aware of its contrary proposition or their doxastic response towards it. But undergoing this process will be necessary if acceptance of \(P\) is also intertwined with believing its contrary proposition that not-\(P\).

As is mentioned, a rational agent goes through this process when, for the sake of argument or for any other reason, she takes a belief that not-\(P\) to be true, accepts and then deactivates it to accept its contrary proposition that \(P\). Hence, the advocates of the belief-acceptance distinction may properly use their theory to attribute a belief that not-\(P\) to her. Nonetheless, the belief cannot be attributed to a self-deceiver, an irrational agent, who doesn’t even take a belief that not-\(P\) to be true under any circumstance (e.g. the self-deceived voter would never take this proposition to be true: my favorite candidate is a fraud), let alone accepting and then deactivating it to merely accept its contrary proposition.

This attitude towards belief attribution gives a special first-person authority to a belief-holder that reminds us of \textit{the constitutive account of self-knowledge} according to which the best way to realize whether one believes a proposition is this: “given certain conditions \(C\), \(S\) believes that \(P\) if and only if \(S\) believes that she believes that \(P\)” (Engel, 2010, P. 12). What is important about this view is that only having a capacity of a second-order belief is sufficient to attribute a first-order belief to an agent, and the agent doesn’t
necessarily need to have the occurrent second-order belief. This view embraces the Cartesian features of authority, privileged access and transparency. Due to having access to our own thoughts in a transparent privileged way, we cannot be wrong about our thoughts. Nonetheless, an opponent of mine may argue that this account cannot explain self-deception. In what follows, I will lay out this objection, and then I will show why it doesn’t work.

4.1) The First Objection

Many would argue that the constitutive view is only applicable to rational agents, and cannot account for self-deception, since it guarantees that a self-deceiver cannot be wrong about her beliefs, whereas a self-deceiver is an irrational agent who makes mistakes in these ways: a) A self-deceiver has a false second-order belief that she doesn’t believe that not-\(P\), while actually holding a first-order belief that not-\(P\). b) A self-deceiver may not really believe that \(P\), while having a conscious second-order belief that she believes that \(P\).

I think this objection doesn’t work, since it is consistent to claim that a self-deceiver (on a par with a rational agent) cannot be wrong about her thoughts, and at the same time show why a self-deceiver is an irrational agent. The rationale for this claim is that what makes a self-deceiver intellectually dishonest is that a self-deceiver has formed her first-order belief based on an irrelevant ground (namely, ignoring relevant evidence or considering irrelevant evidence), and then holds a second-order belief on that wrong basis too. I don’t think that, as the above objection claims, the irrationality comes from holding a conscious second-order belief in contrast with an agent’s first-order belief. I think a self-deceiver is irrational, since she is the same as a biased agent who sees evidence against \(P\) but ignores it (which is an intellectually dishonest act) not to acquire a first-order belief
that not-\(P\), and then builds her second-order belief in accordance with it (i.e. I believe that I don’t believe not-\(P\)). Instead, by focusing on irrelevant evidence (which is another case of an intellectually dishonest act), a self-deceiver tries to obtain a first-order belief that \(P\), and accordingly develops her second-order belief based on her first-order belief that \(P\).

In addition to the above reason, there is also another reason why this objection doesn’t work. As I argued earlier, it is difficult to prove that a self-deceiver necessary holds a first-order belief that not-\(P\). So, the first part of this objection (i.e. a claim according to which \textit{a self-deceiver holds a first-order belief that not-\(P\), without holding a second-order belief in accordance with it}) is problematic. And in the final section, I will show in detail why the second part of the objection (i.e. a claim which states that \textit{a self-deceiver may not really believe that \(P\), whereas her second-order belief leads her to believe differently}) doesn’t work either, since there is enough reason to attribute a belief that \(P\) to a self-deceiver.

4.2) The Second Objection

According to this objection, my criterion of belief-attribution (i.e. taking a proposition to be true) is not broad enough, and so is not the best criterion. This is the line of reasoning Cohen chooses. According to Cohen, to believe that not-\(P\) one must \textit{be disposed to feel} that it is true that \(P\) and false that not-\(P\) without being required to act, reason or speak accordingly. Hence, on this view, it is legitimate to attribute an unpleasant belief that not-\(P\) to a self-deceiver, since a self-deceiver is \textit{disposed to feel} that it is true that not-\(P\) and false that \(P\).
I contend that there is a problem with this sort of belief attribution, since its condition for belief-attribution is so broad that may allow for contrary propositions to be an agent’s belief at the same time. If all we need to attribute a belief that not-\(P\) to a self-deceiver is being disposed to feel that the belief is true and its contrary belief is false, then why not use the same strategy to attribute a belief that \(P\) to the agent too? By Cohen’s logic, attributing a belief that \(P\) to a self-deceiver is even more likely than a belief that not-\(P\), since a self-deceiver is more inclined to value her feeling that a pleasant proposition that \(P\) is true than the other way.

I think many would agree that, for instance, it is even more plausible for the self-deceived voter to be disposed to feel that it is true that \(P\) (i.e. “my favorite candidate is honest”) than not-\(P\). However, if Cohen doesn’t want to deem \(P\) to be a self-deceiver’s actual belief (\(P\) being an actual belief is a more likely conclusion resulting from Cohen’s view), then at least he must admit that his criterion of belief-attribution allows a self-deceiver to show the same doxastic response towards a belief that \(P\) too. This means that, unlike Cohen’s original claim, a belief that \(P\) is also a self-deceiver’s actual belief. This conclusion, on its own, will lead to an impossible state of mind in which we attribute two contradictory beliefs to an agent simultaneously. Cohen might answer that a self-deceiver is not disposed to feel that both \(P\) and not-\(P\) are true simultaneously. As a result, there won’t be any contradiction if a self-deceiver demonstrates the dispositions at issue at different times.

I believe that this potential response encounters the same problem as Donald Davidson’s view (1985). In terms of Davidson’s view, we may attribute two contradictory beliefs to a self-deceiver at different times. But these types of attitudes towards self-
deception are problematic, since they reduce the puzzle of self-deception to a mere problem of simply holding two contradictory beliefs when one belief is standing and another one is occurrent. These views cannot show self-deception’s tension, since in order for the tension to be present, a self-deceiver must have conscious awareness of the two sides of a conflict. Furthermore, it is not determined how these two contradictory beliefs interact with each other when a self-deceiver’s mind is partitioned between two parts.

Another problem for Cohen’s view is that “disposition to feel that \( P \) is true” represents more one’s suspicion, supposition, and/or hope towards something than believing it. Feeling something may give one prima facie justification for believing it, but it cannot necessarily result in ultima facie justification for belief. “Disposition to feel that \( P \) is true” is a broad condition that allows belief-like states (i.e. hope, supposition, suspicion, etc.) to be regarded as belief. For instance, my feeling that “the person on the other side of the street is my father” merely shows my guess that such a thing is the case. But I cannot necessarily claim that I believe that “the person on the other side of the street is my father”, because believing gives rise to confidence that justifies further actions (e.g. calling that person or smiling at him), whereas belief-like states lack confidence, and this is why no one speaks, reasons, or acts merely in terms of hope or suspicion.

4.3) The Third Objection

The advocates of the belief-acceptance distinction would argue that it’s true that a self-deceiver doesn’t verbally take a belief that not-\( P \) to be true. But the agent occasionally acts in a way which makes us justified attributing the belief to her. According to this response, a self-deceiver’s belief can be seen better through her actions than her verbal assertion, since “actions talk louder”, due to occurring more spontaneously and genuinely. For
instance, a self-deceiver might deny verbally that his wife has been having an affair, but spontaneously acts in a way which gives us a reason to attribute this belief to him: “I have been cheated on”. In terms of this view, due to different reasons (e.g. lack of self-knowledge, thinking highly of ourselves, being afraid of confronting the truth, etc.) people may fail to recognize their actual belief. Hence, attributing a belief to an agent based on *an agent’s own recognition of her mental state* is a mistake, whereas a second or third-party tends to be less biased when it comes to recognizing an agent’s belief.

As is mentioned, philosophers like Audi and Cohen follow this path by exploring the existence of *unconscious belief* (i.e. a type of belief which is not occurrent and is represented mostly through one’s actions). According to these philosophers, an agent’s actual belief tends to be represented more genuinely through her behavior, despite the agent’s verbal denial of what her behavior indicates.

From my perspective, there is a problem with this response due to relying on behaviorism. It’s true that owing to different reasons (i.e. weakness of will, phobia, etc.), we may fail to act in terms of our honest verbal assertion or our behavior might contradict our verbal statement. But this doesn’t mean that a second or third-party has a sufficient reason to attribute a particular belief to us only in terms of our behavior. For instance, it is true that individuals with lack of self-knowledge may fail to see a contradiction between their verbal assertion and non-verbal behavior, but even in these cases an agent’s own idea about her internal state is required for distinguishing her actual belief. Take this case to see why an agent’s own idea matters: suppose that a woman sincerely asserts that parents shouldn’t favor one child to another. However, due to an old habit and lack of enough self-reflection, the woman’s behavior contradicts her assertion, while she is not aware of her
behavior’s indication. To put it differently, the reason for that discriminatory act comes from a mere habit (probably common in her childhood environment) which has failed to change in accordance with the woman’s current sincere assertion. But the act’s reason doesn’t derive from the woman’s genuine belief that “favoritism is acceptable”. As a result, it is not fair to attribute the belief at issue to the woman owing to representing a certain behavior indicating so\(^9\). The same goes for a self-deceiver. A second or third-party cannot necessarily attribute a belief that not-\(P\) to a self-deceiver due to a self-deceiver’s behavior.

4.4) The Fourth Objection

There is also another support in favor of the claim that a self-deceiver’s report of her mental state is not reliable. According to this potential objection, a self-deceiver at most can express what she thinks, but she may not legitimately describe her state of mind. This objection relies on Wittgenstein’s distinction between belief expression and belief description. Belief expression usually occurs when one asserts, for instance, sentences such as “I believe that \(P\)” . But sentences of this kind are not necessarily description of one’s state of belief. Belief description occurs mostly with the help of a second or third party in sentences such as this: “the Earth orbits the Sun, but she doesn’t believe it”. In these cases, a second or third party only describes a first person’s state of belief without expressing anything.

\(^9\) Tamar Szabo Gendler (2008) coins the word “alief” to explain the tension between one’s spontaneous behavior and one’s explicit assertion. In terms of this view, for instance, rational tourists may believe that the glass skyway over the Grand Canyon is safe, but alieve that they are in danger which means that they find themselves acting in fear, behaving as though they do not believe it is safe.
The implication of this view for self-deception is that a self-deceiver cannot be the one who correctly describes her own state of mind. The followers of the belief-acceptance distinction may use this theory to argue for their claim: a self-deceiver at most *expresses* a belief that \( P \) and a disbelief that not-\( P \). But she cannot properly *describe* her state of belief. Hence, we need to rely on a second or third party’s reliable report to describe a self-deceiver’s belief which is nothing but believing that not-\( P \) and accepting a fact or a proposition that \( P \) without believing it.

I think an appeal to Wittgenstein’s distinction between belief-expression and belief-description in an above-mentioned way doesn’t seem to be a plausible account of self-deception, since what this view advances is a *normative claim* that suggests what an agent’s ideal belief should be, and so it only works for describing a rational agent’s state of belief who meets her epistemic duty. But it is not determined whether using Wittgenstein’s view in this way can successfully describe a self-deceiver’s *actual* belief. If my opponents’ reading of Wittgenstein’s view is going to work for self-deception, then there won’t be any intellectually dishonest agent (e.g. a biased person, a self-deceiver, etc.), since by assigning to these agents a right belief (that they should have taken ideally), we will rationalize their intellectual dishonesty, and they will not possess any false belief which is a strange result.

I suggest a better way to use Wittgenstein’s view to successfully explain a self-deceiver’s doxastic state. In terms of my interpretation of Wittgenstein’s view, it is still a second or third party who is going to describe a self-deceiver’s state of mind. But this description is more similar to this: “a fact is (or is not) the case, but a self-deceiver doesn’t believe it”. For instance, “a self-deceived husband is cheated on, but he doesn’t believe it”.
This description is closely aligned with what we see in the case of biased agents who deny the truth. And it is also in accordance with my claim throughout the whole paper (i.e. a self-deceiver doesn’t believe an unpleasant justified fact). Instead of using the previous explanation that assigns an ideal true belief to a self-deceiver, an irrational agent, the current view shows a more plausible account of a self-deceiver’s doxastic state.

In this section, I argued that the belief-acceptance distinction cannot successfully attribute a belief that not-$P$ to a self-deceiver. In what follows, by analyzing different features of acceptance, I will show that there is no way for the advocates of this distinction to prove the second part of their claim either (i.e. a self-deceiver only accepts a pleasant unjustified proposition that $P$ without believing it).

5) The Belief-Acceptance Distinction and its Failure to Attribute Mere Acceptance of a Belief that not-$P$ to a Self-Deceiver

I aim to show that different views of the belief-acceptance distinction cannot successfully attribute to a self-deceiver mere acceptance of a proposition that $P$ without believing it. In other words, I am going to demonstrate that in self-deception’s case acceptance of $P$ is always intertwined with holding a belief that $P$. And this is what differentiates my view from the proponents of the belief-acceptance distinction.

I argue for my claim by explaining first when it is legitimate to attribute mere acceptance of a proposition to an agent without assigning its relevant belief to her. I do this through analyzing what a rational agent does: when a rational agent (e.g. a scientist) decides to merely accept something without believing it, her decision occurs voluntarily and at will. The same as any other voluntary action, this voluntary decision requires an
agent’s consciousness of her action. That is, this decision requires an agent to be conscious that or to acknowledge at least tacitly that her acceptance doesn’t have any belief attached to it. An agent here forms the second-order belief that “I don’t believe what I have accepted”. But this is not what a self-deceiver does.

I showed earlier how, for instance, a scientist is conscious or acknowledges that she merely accepts a proposition for the sake of argument without necessarily believing it, whereas a self-deceiver (due to practical or prudential reasons) would never hold the second-order belief that she has only accepted a pleasant fact that \( P \) without believing it. If a self-deceiver admits this to herself, then the fundamental component of deception will be ruled out from the notion of self-deception, and this will lead to the denial of the whole phenomenon of self-deception which is problematic.

Another reason why mere acceptance of \( P \) without believing that \( P \) cannot be attributed to a self-deceiver is that mere acceptance deals with voluntary acceptance which occurs without belief. Voluntary acceptance is different than non-voluntary acceptance which is a continuation of belief. Voluntary acceptance of \( P \) without believing that \( P \) cannot be the case in self-deception, since a self-deceiver is forced primarily by her motives and emotions to hold a belief that \( P \) first. And acceptance of \( P \), as a continuation of a belief that \( P \), appears only after holding the belief in question. Since this acceptance is nonvoluntary, it is always accompanied by its corresponding belief, and so cannot exist by itself.

There is also another reason in favor of my claim that in self-deception, a belief that \( P \) is also attached to acceptance of \( P \). This reason also accounts for a self-deceiver’s intellectual dishonesty, and shows how a self-deceiver and a rational agent differ. What a rational agent (e.g. a scientist) does is not intellectually dishonest, since she only accepts
an unjustified fact *without believing it*, whereas a self-deceiver is irrational, since she *believes* what she *only has a reason to (at most) accept*.

I contend that even if different views of the belief-acceptance distinction find a way to successfully attribute to a self-deceiver holding a belief that not-\(P\) and also mere acceptance of a belief that \(P\), they will encounter another problem again: by attributing the belief and acceptance in question to a self-deceiver, they will automatically rationalize an irrational agent’s epistemic behavior or reduce her to a rational agent. And this sounds counterintuitive. Different views of the belief-acceptance distinction cannot use the same account (which they use to show how belief and acceptance can *rationally* go in a different direction) to explain why a self-deceiver is in an irrational state.

5.1) Objection

An opponent of mine might argue that in the above case at most there exists a belief but there is no acceptance. According to this objection, non-voluntary acceptance is nothing but a belief, since beliefs are held nonvoluntarily, and so it is redundant to come up with a new phenomenon named “nonvoluntary acceptance” when this so-called phenomenon is reducible to the category of belief.

I think this objection is not valid, since, as I mentioned, in order to hold a belief, one needs to take it to be true, whereas acceptance deals with postulating, acting and/ or reasoning in terms of a belief, regardless of whether the belief is considered to be true or not. If one doesn’t deem that a belief is true, but decides to postulate or reason in terms of it, then this is a voluntary acceptance. And if one performs these kinds of behavior as a result of taking a belief to be true, then acceptance occurs nonvoluntarily.
David Clarke (2000) also makes a similar objection when he states that cases of self-deception might be cases of belief in a different proposition that we wrongly assume to be acceptance. I think Clarke is right in his claim that in these cases a self-deceiver holds a belief. But in contrast with him, I think a self-deceiver also does accept a belief that $P$, since she postulates, acts or reasons in terms of it. However, as I mentioned, this kind of acceptance, occurring after holding a particular belief, is different than voluntary acceptance of a fact or a proposition relying on no belief. In order for the latter type of acceptance to be formed, one’s voluntary decision is required, whereas the former roots in one’s belief acquired nonvoluntarily.

Clarke’s view fails to show the compatibility of belief with acceptance, since his account doesn’t consider the distinction between voluntary and non-voluntary acceptance. On Clarke’s view, acceptance is simply a matter of doing something without having any cognitive attitude towards it. The problem with this account is that it only leans towards a voluntary type of acceptance according to which an agent reasons or acts in terms of something, while being aware that these gestures are only an act devoid of any belief.

My explanation properly distinguishes two phenomena from each other (i.e. voluntary vs. non-voluntary acceptance) that their difference has been largely unexplored in the literature. Most of the literature on the belief-acceptance distinction take acceptance to be voluntary without considering the possibility of non-voluntary acceptance as an independent phenomenon resulting from belief.
Conclusion:

I argued that different views of the belief-acceptance distinction cannot successfully solve the puzzle of self-deception, since by attributing a belief that not-\(P\) to a self-deceiver and acceptance of a proposition that \(P\), they treat a self-deceiver the same as a rational agent, and so they fail to capture the irrationality of the phenomenon.

To distinguish a self-deceiver from a rational agent, Cohen refers to a self-deceiver’s mistreatment of evidence (i.e. ignoring relevant evidence and relying on irrelevant evidence). But I demonstrated that even this suggestion cannot differentiate a self-deceiver from a rational agent, since this is what a rational agent does when for the sake of argument, she treats evidence in a special way. I argue that the two agents differ in terms of their belief attitude towards a particular proposition. For instance, a rational agent forms a first-order belief that not-\(P\) by taking a proposition that not-\(P\) to be true, and also holds a second-order belief that “I (still) believe that not-\(P\)” when for the sake of argument, she has to act as if she doesn’t believe that not-\(P\). But a self-deceiver would never take a proposition that not-\(P\) to be true under any circumstance. If this explanation is correct, then there is no way to attribute holding a belief that not-\(P\) to a self-deceiver.

In addition, I also showed that we cannot attribute to a self-deceiver acceptance of a proposition that \(P\) without its relevant belief that \(P\). The reason is that acceptance of a proposition without its relevant belief occurs voluntarily and also requires an agent’s conscious acknowledgement that there is no belief attached to the mere acceptance. But this is not what happens in self-deception, since a self-deceiver would never dissociate herself from holding a belief that \(P\). The agent here is forced primarily by her motives and emotions to form a belief that \(P\), and then acceptance of \(P\) occurs as a continuation of a
belief that \( P \) which is a non-voluntary process. I concluded that the best approach is treating a self-deceiver the same as a biased agent, since she believes what she only has a reason to (at most) accept, and doesn’t believe what she has enough reasons to believe.
Third Paper

A New Version of the Emotion-Based Account of Epistemic Blame

1) Introduction

In *Blame and Wrongdoing*, Jessica Brown (2017) states that there is considerable overlap between ethics and epistemology. Blame is one of those notions that some epistemologists think coexist both in ethics and epistemology. However, in spite of a wide range of theories about the nature of moral blame, talk of epistemic blame in contemporary epistemology is a fairly recent phenomenon.

While ethicists take the notion of blame to mostly diagnose morally wrong behavior, epistemologists show us that agents may be blameworthy due to bad reasoning, relying on irrelevant evidence, etc. What epistemologists try to show is that we blame a wishful thinker for interpreting events in her own favorite way and we blame an individual for relying on a clairvoyant’s words in the absence of any justified reason.

Here is my plan for this paper: I will begin section two by explaining what a target of epistemic blame is. After introducing the two views on this topic, I will demonstrate which view is more plausible. Then, in section three, I will give an overview of different accounts of epistemic blame and identify a problem with each of them. As Cameron Boult (2021) states, there are at least four theories of epistemic blame: a) the agency-cultivation view which is engaged with analyzing the influence of blame on a blamed agent; b) the relationship-based view that deals with the influence of blame on a blaming agent; c) the desire-based view according to which when we blame someone, we believe that the agent holds *a belief badly* (i.e. a term which describes a violation of an epistemic norm culpably),
and we desire that she doesn’t hold the belief; and d) the emotion-based view which focuses on our emotional reactive attitudes of indignation and resentment to epistemic failures.

Moving on to section four, I argue that the best general theory of blame is the emotion-based theory of blame. In section five, the final section, I will formulate this view’s epistemic extension, and explain why this view of epistemic blame has better explanatory power than other theories.

2) What Is a Target of Epistemic Blame?

Some philosophers (Brown 2018; Rettler 2017; Woudenburg, 2009; Nottelmann, 2007) argue that epistemic blame must only target doxastic states (namely, belief, disbelief, and suspension of belief) that are devoid of a violation of moral and practical norms. Nonetheless, as Cameron Boult (2021) states, there may be cases of epistemic blame that don’t directly target one’s doxastic states. So, one can epistemically blame, for example, assertion, an action of inquiry and a particular inference someone makes.

I posit that this view is advantageous, since it covers a broader range of epistemic mistakes that may not occur in a doxastic state. However, this view also allows for beliefs and other doxastic states to be a target of epistemic blame (if they are held in violation of an epistemic norm), since these types of doxastic states are a product of bad reasoning or other epistemic mistakes, and accordingly they will be regarded as a target of epistemic blame as well. Throughout this paper, I will refer to this account of a target of epistemic blame.

To see how a target of epistemic blame may exist independently of moral and practical blame, take these two cases: your friend, without any justification, believes that
“the number 13 brings about bad luck”. And in another example, your friend has a terminal illness, and her husband has cheated on her. Realizing this fact may lead to her death. Due to the big cost, your friend deceives herself into believing that “her husband hasn’t cheated on her”. She does that by relying on irrelevant pieces of evidence and ignoring relevant ones. Holding this belief makes your friend feel better, and this will improve her physical condition accordingly.

In both examples, by holding these beliefs, your friend doesn’t necessarily wrong anyone morally or practically. Moral wrongdoing tends to occur when someone’s right is culpably violated by another person. But these cases are not about harm a person may impose on someone else. Both of these examples center around one’s own private life.

Furthermore, no one is necessarily practically wronged here either, since practical wrongdoing happens when one does something harmful to oneself. In the cheating case, at least in the eyes of some, not only is your friend not practically blameworthy, but she acts in favor of her prudential benefit (i.e., by telling a lie to herself, she will feel better, and this will cure her illness faster as well). And in the number case, even if there may not be any practical benefit for the agent by holding on to her superstitious belief, no practical harm would necessarily threaten her either. But in both cases, the agent has been intellectually dishonest, and we blame her for her purely epistemic mistakes. I offered these examples to show how a target of epistemic blame may be devoid of any moral and practical blame.

The above examples also show that epistemic mistakes may occur in situations with low or high stakes. To put it differently, sometimes a target of epistemic blame is one’s attitude that hardly influences one’s action (e.g. the number case), and sometimes epistemic
blame targets instances of epistemic failures in situations in which one’s practical benefits are in danger (e.g. the cheating case). But, as we saw, in both cases, we blame an agent for purely epistemic mistakes. The difference is that your friend’s holding the belief about the number 13 doesn’t necessarily affect your friend’s practical life. Or at least compared to your friend’s belief about her husband’s faithfulness, her belief about the number 13 may not bring about any significant practical harm or benefit to her. So this is not a situation with high stakes. But in the case of cheating, your friend faces a situation with high practical stakes.

The significance of the above debate for the topic of epistemic blame is that it illuminates how epistemic blame is related to actions (this is noticeable specially in the cheating case). We tend to associate moral blame with actions and epistemic blame with beliefs or attitudes that don’t necessarily lead to actions. But the above explanation clarifies how a target of epistemic blame may give rise to actions as well.

Having specified what a target of epistemic blame is, I will examine different accounts of epistemic blame in what follows and identify an issue with each of them.

3) Different Views of Epistemic Blame

Now the question is what epistemic blame consists in? Different accounts of epistemic blame have offered different answers to this question. In this section, I will summarize the theories and then show why they are problematic. Then, at the end of the section, I will choose a more plausible theory, and in the next sections, I will revise and introduce it as the best theory of epistemic blame.
Different views of epistemic blame may be classified into two general groups depending upon whether they analyze epistemic blame with regard to its consequence or its nature. I am going to begin with the former:

a) The first account is called the agency-cultivation view which analyzes epistemic blame in terms of its consequences for a blamed agent. Adam Piovarchy (forthcoming) is a main advocate of this view, and it is on a par with “forward-looking approaches to moral responsibility” (Vargas, 2013). Piovarchy argues that when we epistemically blame someone (by verbal condemnation, calls for shame, emotional distance, etc.), we discourage the person’s exhibiting certain epistemic behavior, and this will cultivate the blamed agent’s epistemic agency.

The problem with this view is about failing to account for cases in which blame occurs without leaving any consequence for a blamed agent. We know that there are cases in which we morally blame, for instance, a psychopath for killing people by showing how repulsed we are by her act of killing people while, due to the psychopath’s flawed character, she fails to cultivate her moral agency. Likewise, we may blame a person for her epistemic mistake, whereas the person is incapable of cultivating her character due to, for instance, being raised in an epistemically biased community. So, in this case epistemic blame still occurs, although it is not accompanied by any agency cultivation. But according to the agency-cultivation view, due to a lack of any cultivation here, the concept of blame is not applicable to these cases.

I think the agency-cultivation view faces this problem, since it wrongly assumes that it is a blamed agent who has to carry the weight of blame. It is worth noting that even in the case of self-blame (i.e. when a blaming agent and a blamed agent are the same),
blame occurs when a blamed agent puts herself *in the position of a blaming agent* and feels guilt or shame as a result of blaming herself. My point here is that there is a blaming agent whenever blame occurs, and this blaming agent can be a person who blames others or blames herself. To put it differently, in order for blame to happen, there should be a blaming agent (whether this person is going to be the one who blames others or blames herself). But the agency-cultivation view cannot show this, since on this view blame isn’t appropriate as long as a blamed agent fails to cultivate her character, even if there is a blaming agent (somewhere else) who blames the person.

b) The relationship-based view is the second theory which defines epistemic blame by focusing on a consequence of epistemic blame for a blaming agent. The relationship-based view is inspired by the works of T. M. Scanlon (2008; 2013) on moral blame. Boult (2020; 2021), a main proponent of this theory of epistemic blame, argues that we are members of an epistemic community, and we stand in “an epistemic relationship” with one another. Standing in this epistemic relationship gives rise to certain epistemic expectations and demands directed towards our epistemic agency. According to this view, we modify our intention to *epistemically trust* those who fall short of these epistemic expectations. And this is what epistemic blame consists of.

Although this view centers around a blaming agent’s attitude to analyze epistemic blame, it still encounters a problem similar to the problem with the previous view. In general, the problem with the relationship-based view (whether in its moral or epistemic version) is that modifying our intention to morally/epistemically trust those who fail to meet moral/epistemic expectations is merely a consequence of blame which is not always the case in all instances of blame. In other words, we may blame someone for her mistake,
while due to different reasons (e.g. suffering from weakness of will, being close to a blamed person, and so hoping to see improvement in her behavior, etc.), we may fail to change our intention to trust her morally/epistemically. Or we may do that but in the long run and in further instances of moral/epistemic blame but not quickly after one instance of mistake.

It seems that losing epistemic trust depends upon a mistake’s severity and/or a blaming agent’s epistemic taste. If I am right, then the advocates of this view should answer why they appeal to a condition which varies from person to person or case by case. This shows that *modification of our intention to epistemically trust* a blamed agent cannot be essential to the definition of epistemic blame.

As is stated earlier, the above-mentioned views of epistemic blame are engaged with a consequence of blame. In the following, I will introduce two views that analyze blame with regard to its nature.

c) The desire-based view, advanced by Jessica Brown (2019), defines the concept of epistemic blame by relying on a blaming agent’s desire. This view is inspired by George Sher’s (2006) view of moral blame according to which moral blame occurs when a blaming agent shows a characteristic set of dispositions (e.g. dispositions to reproach, feeling angry, requesting a reason, or apologizing in the case of self-blame\(^\text{10}\)) which is causally connected to a specific belief-desire pair: the belief is that an agent *has acted badly* which means that

\(^{10}\) One might object that former instances in the parentheses are a *blaming agent*’s reaction, while the latter one (apologizing in the case of self-blame) is a disposition attributable to a *blamed agent*. I think the answer to this objection is that in the case of self-blame, a blamed agent also plays the role of a blaming agent by blaming herself. That is, in the case of self-blame, any blamed agent would also be a blaming agent. Thus, apologizing may be identified as a blaming agent’s reaction in this sense.
the agent has violated a moral norm culpably, and the desire is that the agent has not acted badly.

Brown appeals to this view to make an analogy between moral and epistemic blame: epistemic blame happens when someone, including oneself, believes badly (i.e. believing badly, according to Brown, refers to forming a belief by violating an epistemic norm without having any excuse such as strong but misleading evidence), and our desire that the person doesn’t believe badly is frustrated (this could be a weak version of desire). Brown refers to Bonjour’s familiar case of Maud which is as an example of blameworthy belief: Maud relies on her clairvoyant power (without having any reliable evidence for the existence of this power) to conclude that the President is in New York, while possessing strong reliable evidence that the President is not in New York. Brown states that it is far from obvious that there is a violation of moral or professional principles in this case. So, if we want to blame Maud for her belief in this example, our reaction will be an instance of epistemic blame which occurs in this way: we believe that Maud has a belief badly, and we desire that she doesn’t hold the belief (Brown, 2018, P. 3)

This view doesn’t assume that all agents all the time strongly desire that no one, including themselves, believes badly. Brown acknowledges that there are cases in which individuals pass up opportunities to give others true beliefs or correct their false beliefs even if this doesn’t cost them a lot. She is also aware that there are even cases in which we intentionally want someone (e.g. our enemy) to hold a belief badly without having any frustrated desire about it. Due to the lack of desire in these cases, Brown refuses to consider them as an instance of epistemic blame. According to Brown, her view doesn’t imply that we should always form the desire in question whenever there is a bad belief somewhere.
As Brown states, she merely aims to *describe* what happens when we epistemically blame someone for a *particular belief*.

I contend that the desire-based view is problematic, since it fails to account for those cases in which the phenomenon of blame exists without the belief-desire pair. For instance, take an *akratic* agent (i.e. a person with weakness of will) who believes that smoking is bad, and blames herself for doing that, however, desires to smoke and acts upon her desire. The desire-based view fails to identify the *akratic* agent’s attitude towards her behavior under the title of blame even if the person sincerely judges that “she should not have smoked” and feels ashamed because of this, while at the same time (due to her weakness of will) cannot help desiring to smoke.

The advocates of the desire-based view don’t comprehend that *akratic* agents may sincerely believe that “Φ should not have happened” or “whoever does Φ is blameworthy”, and they may feel guilt or shame after doing Φ, while their judgment and emotion fail to influence their action. In these cases, what motivates the agents’ action is not their belief and emotion, but their desire to Φ. However, the fact that the agents hold the desire at issue doesn’t indicate that the agents don’t find their behavior blameworthy, whereas, according to the desire-based view, these are not instances of blame, since the so-called desire not to Φ doesn’t exists here. The question for the advocates of the desire-based view is: “what else these cases could be called if they are not an instance of blame?”

According to the emotion-based view, we blame or even resent others for forming faulty or hasty belief about significant matters or when they fail to believe when they should (McHugh, 2016, p 66). As Boult mentions, there are only a few philosophers who have spent time discussing the emotion-based view explicitly. However, this view is
widely presupposed in the literature and discussions in epistemology. For example, Nikolaj Nottelmann states that it’s highly plausible to form the feeling of indignation or resentment (at least mild degrees of these attitudes) towards those who are prone to be blamed epistemically (Nottelmann, 2007, p. 3). Lindsay Rettler also presents three paradigm cases of epistemic blame, and then states that blaming agents hold others responsible for their faulty belief by feeling resentment, indignation, and guilt, respectively (Rettler, 2017, P. 4).

I contend that this version of the emotion-based account is also problematic. As the advocates of this view also acknowledge, we tend to express anger or resentment towards those epistemic mistakes which are about significant matters. But the point is that most of the time the significant matters, that noticing a mistake about makes us angry or resentful, are not purely epistemic. My point here is not that purely epistemic topics cannot be significant. All I am arguing for is that when it comes to expressing anger or resentment towards someone for her epistemic fault, this anger or resentment tends to target those topics that matter to us, and most of the time the reason why the topics matter to us is that they are tied to our moral or practical concerns. In other words, the reason is more non-epistemic than purely epistemic. So we express anger or resentment towards those who make epistemic mistakes about these issues, since these mistakes may threaten our moral or practical life. I think it is fair to say that first and foremost what is a main target of our anger or resentment is a violation of our moral or practical norms, and then everything that is an instance of this violation (this could be everything such as epistemic failures, aesthetic mistakes, etc.), is the second target of our anger or resentment. Take this case: you blame a fortuneteller by resenting her for spreading out her superstitious beliefs about the existence of aliens and their impact on human life. You do that since this belief will leave
a bad effect on the youth. You also blame the youth who hold the unjustified beliefs\textsuperscript{11}. Nonetheless, you don’t necessarily resent the latter, since they are victims themselves, and anger or resentment targets someone who is responsible for corrupting the youth, but not someone who is victimized.

As Boult (2021) also argues, feeling indignation or resentment is a more appropriate response when we hold someone morally blameworthy. I think the reason is similar to what I mentioned earlier: moral issues are more related to our life than purely epistemic topics. Hence, a moral mistake threatens one’s practical or prudential concerns more, and this is why more anger or resentment is more appropriate for morally wrong behavior than for a mistake about a purely epistemic topic.

If I am right, then \textit{this version} of the emotion-based view cannot define epistemic blame as an independent category, since, according to this version, epistemic blame is tied to indignation or resentment. But, as is shown, these emotional reactions tend to appear more when our moral or practical benefit is threatened. And, as I explained, even if we express anger or resentment towards purely epistemic mistakes, this is mostly because our moral or practical values are in danger as a result of these mistakes. In other words, the account in question at most may put some epistemic blame in the category of moral or practical blame.

An opponent of mine would argue that this is not a problem for this view. The opponent would conclude from the above example that epistemic blame is not a distinctive kind of blame, since epistemic mistakes matter only when they create a problem for our

\textsuperscript{11} In this case, we may blame the youth by being disappointment with them. In the next sections, I will explain this point more in depth.
moral or practical benefits. In terms of this objection, we hardly care about our own epistemic failures, let alone blame others for their epistemic mistakes especially when our moral or practical values are not in danger because of these mistakes\textsuperscript{12}.

I think this is the wrong conclusion. Of course at times we may value knowledge for its own sake, or we care about something due to purely epistemic reasons. But the point is that when someone makes a mistake about the purely epistemic issues, it is less likely for us to show anger or resentment due to concern for epistemic normativity \textit{per se}. For instance, it is more likely for you to be resentful towards your neighbor’s faulty belief about the number of planets when her holding this belief causes a problem for your daily routines. But you will probably merely be disappointed with her if she holds the belief without leaving any bad influence on your life.

I criticized the emotion-based view of epistemic blame, but this doesn’t mean that there is no way to revise the view in question. In the final section, I will argue for the most plausible theory of epistemic blame by modifying the emotion-based view of epistemic blame. Before that, I need to develop \textit{a paradigm form of blame}. As we saw, all of the above-mentioned theories of epistemic blame are an epistemic extension of a moral theory of blame (which is a foundation of the epistemic theories). But I refuse to refer to the paradigm form of blame as a \textit{moral} theory of blame necessarily, since this kind of labeling rules out practical blame which is another form of blame on a par with moral blame. The

paradigm form of blame should be defined such that different forms of blame are included equally under the title of blame.

4) A Paradigm Form of Blame

In this section, I am going to answer this question: “what do we do when we blame someone?” I will prove that this question’s answer leads us to conclude that the emotion-based account of blame is the most plausible theory of blame in general. After doing this, I will demonstrate how this view is applicable to its epistemic extension.

In order to determine what we do when we blame someone, I refer to a common ground that all the above-mentioned theories of blame adhere to. In terms of these views, the normativity of blaming must be part of any account of blame. For instance, according to the agency-cultivation view, by blaming someone we discourage the person from exhibiting certain epistemic behavior. While on this view the normativity of blaming targets a blamed agent’s behavior, according to the relationship-based view, the normativity at issue encourages a blaming agent to change her behavior towards a blamed agent. These two views have a thing in common: there is a norm that agents (either a blaming agent or a blamed person) should follow in order not to be blamed.

The desire-based view also shows that a blaming agent’s desire informs us about a norm: a certain belief or behavior should not have happened. This view is similar to the emotion-based account according to which a blaming agent’s anger or resentment indicates that her expectations should have been met. Since all of these views endorse the normativity of blaming, I take the normativity at issue as my starting point for explaining the phenomenon of blame.
Since blame always involves a norm, and norms are derived from judgment, blame always involves judgment. And since judgment is a cognitive mental state, blame also inherits the features of judgment, but at the same time blame is more than a cold judgment. For instance, simply judging that “the weather is cold” or “you should not smoke at a hospital” cannot be instances of blame. If blame were to be a mere judgment, then you should be a target of blame when a nurse at a hospital in response to your simple question states that “you should not smoke at a hospital”. But this sounds bizarre, and no one calls this case an example of blame, especially when you don’t even know about the hospital’s policy, and there is no way for you to know that either. So the question is “what else besides a mere judgment is involved in a concept of blame?” This question may be answered by investigating an actual case of blame and analyzing what happens when we blame someone. In what follows, I will assess this case.

In the above example, you are not a target of blame. But we think that you are a target of blame if you keep smoking at the hospital, while you are aware of the prohibition of smoking there. We also think that now the nurse will be justified if she feels anger or resentment towards you. In general, anger or resentment is felt with changes in one’s body when one judges that something unpleasant has occurred. And in this case, the nurse’s anger or resentment occurs, while she judges that you are a target of blame. As a result of this anger or resentment, the nurse may be motivated to take a course of action against you (e.g. informing the security officers, asking you to leave, etc.). And you may also be
motivated to change your behavior due to being ashamed. We may call this second case “the blame case\textsuperscript{13}”.

If the above interpretation of the case is correct, then we must conclude that blame is “a judgmental state accompanied by feelings in a certain way”. This definition shows that blame has two main components: it always comes along with the cognitive mental state of judgment, and also leads to felt changes in the body. Blame may also motivate both agents (i.e. a blaming and a blamed agent) to act in a certain way (e.g. in the second case, the nurse informs the security officers, or you turned off your cigarette). However, the disposition to act in a certain way is not essential to the definition of blame. It is not implausible to imagine, for instance, a case in which we blame someone who mistreats us in public, but we control our anger or resentment in order not to get involved in a fight. Or it is quite plausible to imagine someone who feels ashamed but cannot help acting inappropriately due to suffering from weakness of will.

I think the blame case is compatible with the emotion-based account of blame. Different views of emotions show us how to classify blame as an emotion. According to the cognitive theory of emotions (advanced by Aristotle c, 350 BC 1984: Spinoza 1677/1994: Solomon 1976; Nussbaum 2001), emotions always involve judgments. And the analysis of blame in the second case shows us how blame has roots in judgment.

And for most non-cognitive theorists of emotion (William James 1884; Carl Lange 1885 cited in Prinz, 2009, p. 682), emotions are identified with bodily changes. Felt bodily

\textsuperscript{13} One might object that in the previous section I mentioned that the paradigm case of epistemic blame should be indifferent to moral or practical blame. But the example here has a moral connotation. While the point about the example is correct, it doesn’t create a problem for my claim. I think we may easily replace the content of “the blame case” with another example with a practical connotation.
changes also exist in the case of blame. When we blame others, we feel anger or resentment through our racing heartbeat or breathing quickly. Or in the case of self-blame, we feel ashamed and notice this by our face turning red or not being able to make eye contact.

Here I don’t aim to prove which theory of emotions is most plausible. I am only pointing out that there are sources available to show how blame may qualify as an emotion. If it is right that blame is a type of emotion, then this conclusion will be in favor of the emotion-based account as the most plausible theory of blame, and accordingly the most plausible account of epistemic blame. Nevertheless, as is mentioned, the type of emotion deployed in an epistemic account of blame may differ in a moral theory of blame. I will develop the point later.

The blame case also favors the emotion-based account by ruling out its rival theories (namely, other theories mentioned above). Take the agency-cultivation view and the relationship-based account: these views assume that there is a necessary connection between blame and being motivated to act in a certain way (e.g. according to the agency-cultivation view blame occurs when a blamed agent is discouraged to repeat her blameworthy behavior, and in terms of the relationship-based account blame happens when a blaming agent modifies her intention to epistemically trust a blamed agent). Due to relying on a necessity claim, views of this kind advance some kind of Motivational Internalism (i.e. a view which says that if an agent sincerely judges that it is morally right to $\phi$ in circumstance $C$, then, as a matter of conceptual necessity, she will be to some extent motivated to $\phi$ in $C$). Since the advocates of these views defend a necessity relation here, they fail to account for those cases in which a phenomenon exists without the so-called necessity connection between the phenomenon and one’s motivation to act. But what
occurs in the blame case shows that one’s disposition to act is not necessarily embedded in
the phenomenon of blame, and this is what the relationship-based view and the agency-
cultivation account cannot demonstrate.

In addition, the agency-cultivation view faces another problem that the emotion-
based view doesn’t. As stated earlier, the agency-cultivation requires a blamed agent to
carry the weight of blame such that the phenomenon of blame will not occur if a blamed
agent is not encouraged to change her behavior. But this will lead to a strange result: by
this logic we must accept that anger or resentment a woman (whose son has been murdered
by a psychopath) expresses towards a psychopath is not blame, since the psychopath fails
to change her demeanor. The emotion-based view doesn’t face this problem, since this
view treats blame as a mental state a blaming agent possesses, regardless of whether blame
will leave any influence on a blamed agent or not.

The agency-cultivation view wrongly treats the phenomenon of blame the same as
the act of understanding. Understanding won’t occur if a person (who is supposed to
understand) doesn’t realize a point. But the same cannot be said about emotions (e.g. anger,
resentment, disappointment, etc.) that their occurrence solely depends upon how an angry
or a disappointed person feels. For these emotions to occur, it doesn’t matter how a person,
who is a target of these feelings, reacts. The blame case shows that the emotion-based view
is immune to this problem, since this view classifies blame as a type of emotion a blaming
agent experiences, whereas the agency-cultivation view relies on a blamed agent to create
the phenomenon of blame.

Now I am going to explain how the blame case rules out the desire-based view as
well. As I mentioned earlier, the desire-based view fails to account for those cases in which
we may refuse to blame someone in spite of believing that the person acts or believes badly and desiring that such a thing doesn’t happen. I posit that since blame is an emotion, it goes beyond one’s beliefs and desires, and owing to this reason, the desire-based view doesn’t have any explanatory power here. Suppose that in the blame case, the nurse believes that you have acted badly by smoking, and she also desires that you haven’t done that. However, she is aware that smoking is the only way by which you reduce your anxiety in that situation, and consequently, she doesn’t experience anger or resentment towards you which indicates that she doesn’t blame you. But according to the desire-based view, blame occurs whenever a certain belief-desire pair exists.

I am moving on to a new objection: A proponent of the desire-based account may argue that emotions such as anger, resentment, and disappointment (in the case of epistemic blame\textsuperscript{14}) are a result of a failed desire themselves, and so the emotion-based view is reducible to the desire-based account. According to this objection, for instance, an emotion such as disappointment occurs when one feels dissatisfied due to having her expectations, hopes and desires unfulfilled. The emotion of disappointment correspondingly may lead to other emotions such as anger or resentment in severe circumstances. Since a failed desire is a basis of disappointment, then it will be a basis of anger as well. And consequently, there won’t be any difference between the emotion-based account and the desire-based view stemming from desires. If this is true, then we needn’t appeal to the emotion-based view.

\textsuperscript{14} In the next section, I will develop the idea that disappointment may be replaced with anger or resentment in the case of epistemic blame.
I think this objection doesn’t work. It’s true that a failed desire is a foundation of disappointment. But emotions such as disappointment, anger or resentment go beyond a mere desire. Take this analogy to see how emotions and desires may go to separate ways: you may desire to take your medication without necessarily a feeling of joy sweeping over you, whereas in other cases (e.g. drinking a cup of coffee) your desire may be accompanied by the feeling at issue.

In the case of a smoker with anxiety, we saw that you are the smoker who is directly responsible for our failed desire. The emotion-based account makes it possible for us not to express our anger or resentment towards you, despite the fact that we desire that you don’t act in a certain way. The emotion-based view even allows us to extend the target of our anger or resentment to others different than you, those who provoke your anxiety. In contrast, according to the desire-based view, blame occurs upon judging that our desires are failed. Furthermore, the advocates of the desire-based view believe that the target of blame is necessarily someone who is directly responsible for our failed desire.

In sum, if the emotion-based account were reducible to the desire-based view, the followers of the emotion-based view would not be able to withhold responsibility from someone who has not fulfilled our desire, or they would not be able to extend the target of blame to someone different from an agent whom we hold directly responsible for our failed desire.

So far, I have defended the plausibility of the emotion-based account of blame by developing a case with a moral content. But I haven’t applied the theory to epistemic blame yet. In what follows, I am going to show the epistemic extension of the emotion-based account.
5) The Emotion-Based Theory of Epistemic Blame Revised

As is mentioned, the emotion-based view (in its original version) relies on the emotions of indignation or resentment to account for epistemic blame. But we saw that these emotions don’t seem to be an appropriate response for epistemic blame. However, this doesn’t mean that emotions cannot be a solid basis to build a theory of epistemic blame upon. What we need to do is to find the right kind of emotion here. Nonetheless, one might object that appealing to a different type of emotion as a foundation of epistemic blame is not a good strategy, since there won’t be a unified sense of blame anymore, and this will lead to a change in the whole concept of blame.

I think this objection fails, since as it is possible to have different emotions when it comes to blaming others and ourselves morally (namely, the emotion of guilt or shame in the case of self-blame and anger or resentment when we blame others), it is also possible to experience another kind of emotion when it comes to epistemic blame (which is a distinctive form of blame different than moral and practical blame).

Miranda Fricker (2016) argues that there is very little present in all cases of blame. Blame in one context can vary significantly from blame in another context. If this is right, then there is no need to be necessarily committed to the same type of emotion to account for different kinds of blame. Hanna Pickard (2013) takes this line of reasoning when she states that disappointment and discontent can be blame emotions. But Pickard doesn’t

\[15\] To simplify, I only refer to the emotion of disappointment, since there is not a fundamental difference between disappointment and discontent.
develop a theory of epistemic blame based on the emotions in question. And this is what I am going to do.

In the previous paragraphs, I explained what I mean by disappointment (i.e. an emotion which is commonly associated with or represents the feeling of dissatisfaction that appears when one’s expectations and hopes are not fulfilled). In addition, I refuted the objection that emotions such as disappointment are reducible to desires. In what follows, I aim to show how disappointment accounts for epistemic blame. To answer this question, I need to refer to the target of epistemic blame one more time. As I mentioned before, you may blame your friend for her belief that “the number 13 brings about bad luck”.

I stated earlier in this case of epistemic blame for holding these beliefs, your friend doesn’t necessarily wrong anyone morally or practically. As a result, it doesn’t sound plausible to express anger or resentment towards your friend. Anger or resentment is a more plausible response when your friend holds the belief that, for instance, “polygamy is a decent lifestyle” rather than her belief about the number 13. While your friend’s idea about the number 13 may also lead to strange results, your friend’s idea about the permissibility of a polygamous lifestyle is more blameworthy, since it interferes with many people’s moral or practical benefits (e.g. it may corrupt the youth, ruin families, etc.).

In the case of polygamy, our motivation to blame the person comes from care and concern for moral or practical normativity. Thus, in this case, we deal with moral or practical blame. But in the case of number 13, we blame an agent epistemically, since we are motivated by purely epistemic reasons.
Now the question is this: “what is an appropriate response to a type of epistemic mistake that doesn’t violate our practical or moral norms?” We know that our reaction to a blamed agent is different, depending upon the severity of a mistake. Since making a simple epistemic mistake doesn’t necessarily put one’s moral or practical benefits in danger, a more minimal conception of blame would suffice to account for epistemic blame. Hence, a mild negative emotion such as disappointment is a good choice to fulfill the minimal conception of blame here. Appealing to this minimal conception of blame makes it possible for us to classify our reaction to certain epistemic mistakes under the title of epistemic blame without necessarily moving beyond the notion of blame to explain these cases.

In the previous paragraphs, I demonstrated how the emotion-based view is advantageous compared to other accounts of blame. I showed this through investigating a blame case with moral connotations. I think it is also quite possible to refer to one of those cases of epistemic blame (e.g. the cheating case) and prove that the emotion-based account of epistemic blame explains this case better than its rivals.

In the case of cheating, we saw that evidence indicates that your friend’s husband has cheated on her. Your friend has a terminal illness and realizing this fact may lead to her death. Due to the big cost, your self-deceived friend manipulates evidence to believe that “her husband hasn’t cheated on her”. According to the emotion-based view, epistemically speaking, you are disappointed by your friend due to her intellectually dishonest behavior (e.g. relying on misleading evidence culpably to form a false belief). Hence, you find her epistemically blameworthy. However, practically speaking, you desire that the friend holds the belief due to its practical benefit.
The emotion-based view shows how epistemic blame exists in this case. But the desire-based view cannot account for epistemic blame in the above case (in spite of the existence of an epistemic mistake there), since you don’t desire that your friend not hold the belief.

The relationship-based account won’t work here either, if you fail to modify your intention to epistemically trust your friend on matters concerning her husband’s fidelity. But it is imaginable that in this case, your epistemic expectations are not met by your friend, and you are disappointed consequently, however, you don’t know that you must also modify your intention to epistemically trust your friend upon judging that she fell short of your epistemic expectations. Or you may want to modify your intention, but you fail to do that, since she is a close friend of yours, and it is hard for you to give up on her epistemically (even about issues such as her husband’s fidelity). The problem with the relationship-based view is that, in terms of this view, there is no epistemic blame here due to a lack of your modified intention. But, as I stated, epistemic blame exists here by the manifestation of your emotional reactive attitude towards your friend’s epistemic mistake.

And the advocates of the agency-cultivation view would argue that epistemic blame won’t occur if your friend fails to cultivate her epistemic agency (after you discourage her behaving in an epistemically improper way). But the problem with this view is that it doesn’t leave any room for the mental act of blame which may occur only in the mind of a blaming agent. In this case, you blame your friend, although you don’t express it due to certain practical concerns. But the agency-cultivation view cannot explain this type of blame.
The defenders of the agency-cultivation view would reply that since in this case the blaming agent doesn’t even express her disappointment, then there is no way for the blamed agent to know that she is epistemically blameworthy. And owing to the fact that the blamed agent is not aware of this, she will not be encouraged to correct her epistemic mistake. Thus, this case doesn’t create a problem for the agency-cultivation view, since, according to the agency-cultivation view, a blamed agent’s cultivation of her character happens after being discouraged by a blaming agent from repeating her mistake. But this case is created such that its representation of blame is very different than what the agency-cultivation view claims. This case doesn’t even meet the first requirement of blame in terms of the agency-cultivation view: (i.e. a blaming agent’s discouraging a blamed agent from acting in a certain way). Therefore, this case may not be able to rule out the view in question.

I think this defense of the agency-cultivation view fails for these reasons: first, this defense assumes that only internal criticism is legitimate, and since the above case of blame fails to criticize the agency-cultivation view internally, then this is not a valid objection to this view. I contend that there is no reason to believe that internal criticism (at least in this case) is the only legitimate type of critique. Sometimes a view is so flawed that there is no reason to necessarily accept its general outline to criticize it. Second, for the sake of argument, I assume that this defense of the agency-cultivation view is legitimate, and we must design the blame case in a different way. But it is imaginable that, even in that case, after showing your disappointment, your friend would not necessarily cultivate her epistemic agency due to being biased. Cases of this kind may occur a lot when a biased person refuses to change her belief in spite of the existence of evidence. If this is right, then
advocates of the agency-cultivation view will not be able to apply the concept of blame in this case either.

**Conclusion:**

I laid out different theories of epistemic blame and identified a problem with each of them. Then, I chose the emotion-based account of epistemic blame as a view which has the potential to be successfully revised. In order to introduce a new version of the view, I created a paradigm form of blame. The case shows that blame is normally associated with emotional reactions such as anger and resentment towards a target of blame. By appealing to this conclusion, I argued that the emotion-based view is the most comprehensive theory of blame. But resentment and anger are not a suitable emotional response to epistemic failures, since a target of epistemic blame is different than a target of moral and practical blame. Thus, I referred to a variable account of blame according to which blame may be represented through the emotion of disappointment or discontent as well. I applied this view to epistemic blame and suggested that the revised version of the emotion-based view of epistemic blame is the most plausible theory of epistemic blame. Finally, I showed how the view in question has better explanatory power than the other views.
Conclusion:

In my first paper, I argued that Marcus’ view fails to defend the Knowledge Thesis from the puzzle of self-deception, since Marcus’ account of self-deception is problematic. What is important for any account of self-deception is showing that a self-deceiver has conscious awareness of a proposition that not-\( P \) when he sincerely asserts a proposition that \( P \). Conscious awareness is required here to show how a self-deceiver suffers from tension. Otherwise, a self-deceiver will be regarded the same as a person with a lack of self-reflection. But Marcus argues that a self-deceiver possesses masked knowledge to show that a self-deceiver unconsciously knows a proposition that not-\( P \) while non-lyingly asserting a proposition that \( P \). I showed that masked knowledge cannot be unconscious, since one cannot intentionally make one’s conscious knowledge unconscious.

Marcus also argues that even if a self-deceiver is aware that \( y \)-ing will promote \( x \)-ing, he doesn’t see that he is \( y \)-ing intentionally or he doesn’t see that it is for the sake of \( x \)-ing that he is \( y \)-ing. I demonstrated that Marcus’ view is not consistent in attributing knowledge to a self-deceiver. If a self-deceiver knows that he is \( y \)-ing, and also is aware that \( y \)-ing will promote \( x \)-ing, then he cannot consistently claim that he doesn’t see that his \( y \)-ing leads to \( x \)-ing or he doesn’t see that he is \( y \)-ing intentionally.

Having shown why Marcus’ view is flawed, I used the theory of ‘knowledge without belief’ to show the compatibility of self-deception with Knowledge Thesis. The theory in question can successfully account for the component of conscious awareness by attributing knowledge to a self-deceiver without requiring him to necessarily believe that not-\( P \). Since no belief is required here, then we don’t need to resort to masked knowledge or unconscious belief to reconcile holding a belief that not-\( P \) with a belief that \( P \). However, this view allows
us to explain a self-deceiver’s tension by showing that a self-deceiver is conscious of an unpleasant fact (which makes him suspicious that the unpleasant fact is the case) while trying not to believe it or getting himself to believe the other way.

In the second chapter of my dissertation, I argued that different views of the belief-acceptance distinction cannot successfully solve the puzzle of self-deception, since by attributing a belief that not-\(P\) to a self-deceiver and acceptance of a proposition that \(P\), they treat a self-deceiver the same as a rational agent, and so they fail to capture the irrationality of the phenomenon. I also argued that even Cohen’s suggestion (i.e. a self-deceiver’s manipulation of evidence) cannot distinguish a self-deceiver from a rational agent, since the two agents differ in terms of their belief attitude towards a particular proposition. For instance, a rational agent forms a first-order belief that not-\(P\) by taking a proposition that not-\(P\) to be true, and also holds a second-order belief that “I (still) believe that not-\(P\)” when for the sake of argument, she has to act as if she doesn’t believe that not-\(P\). But a self-deceiver would never take a proposition that not-\(P\) to be true under any circumstance.

In addition, I also showed that we cannot attribute to a self-deceiver acceptance of a proposition that \(P\) without its relevant belief that \(P\). The reason is that acceptance of a proposition without its relevant belief occurs voluntarily and also requires an agent’s conscious acknowledgement that there is no belief attached to the mere acceptance. But this is not what happens in self-deception, since a self-deceiver is forced primarily by her motives and emotions to form a belief that \(P\), and then acceptance of \(P\) occurs as a continuation of a belief that \(P\) which is a non-voluntary process. Furthermore, I showed that, unlike a rational agent, a self-deceiver would never dissociate herself from holding a belief that \(P\). A self-deceiver, the same as a biased agent is irrational, since she believes
what she only has a reason to (at most) accept, and doesn’t believe what she has enough reasons to believe.

In my third paper, I criticized different theories of epistemic blame, and picked out the emotion-based account of epistemic blame as a view which has the potential to be successfully revised. In order to introduce a new version of the view, I identified a paradigm form of blame first by analyzing a case. The analysis of the case showed that blame occurs upon judging that an agent is responsible for a culpable violation of a norm. This judgment comes along with emotional reactions such as anger and resentment towards a target of blame. By appealing to this conclusion, I argued for the emotion-based view as the most comprehensive theory of blame in general. But I also argued that resentment and anger are not a suitable emotional response to epistemic failures, since a target of epistemic blame is different than a target of moral and practical blame. Thus, I referred to a variable account according to which blame may be represented through the emotion of disappointment or discontent as well. I applied this view to epistemic blame and suggested that the revision of the emotion-based view of epistemic blame is the most plausible theory of epistemic blame. Finally, I demonstrated that the emotion-based account of epistemic blame has an advantage over other views by having better explanatory power than its rivals.
Bibliography:


University Press.


VITA

Zeinab Rabii was born and raised in Iran. She went to the University of Tehran where she earned two BAs: one in Persian literature and another one in philosophy. She was graduated there in 2009. Then she obtained an MA in philosophy from Allameh Tabataba'i University. In 2015, she went to the University of Wyoming to get an MA in philosophy. While she was about to finish her MA there, she was admitted to the philosophy program at the University of Missouri-Columbia where she got her second MA in philosophy in 2020.