Universidade de Lisboa
Faculdade de Letras

Epistemic Akrasia and Constraints on the Formation of Beliefs
An essay on the characterization and rationality of epistemic akrasia

Mestrado em Filosofia

João Pedro Cesário Miranda

2022

Dissertação especialmente elaborada para a obtenção do grau de Mestre, orientada por Professor Doutor António José Teiga Zilhão

Abstract

Acrasia epistémica é a contraparte teórica de um fenómeno que tem despertado a curiosidade de filósofos com um interesse em raciocínio prático desde Aristóteles. Contudo, quando comparada com a sua contraparte prática, a acrasia epistémica percorreu uma viagem muito mais curta na história da filosofia. Apesar do facto de o debate sobre acrasia epistémica apenas ter começado a ganhar tracção nas últimas décadas, não é frequente (pelo menos, não o suficiente) o recurso ao progresso feito ao longo de milénios sobre o problema da acrasia prática (salvo algumas excepções). Neste ensaio, começarei por fazer um levantamento das principais abordagens à acrasia epistémica disponíveis na literatura, antes de apresentar a minha proposta: que uma teoria que explica a acrasia epistémica ao apelar à maneira como crenças são formadas sobre diferentes tipos de constrangimentos pode ser expandida com sucesso para explicar casos de acrasia epistémica com resultados igualmente bem sucedidos. Terminarei mostrando como esta proposta pode ser integrado no programa da racionalidade limitada, providenciado, portanto, um enquadramento a partir do qual a racionalidade de casos de acrasia epistémica pode ser avaliada.

Palavras-chave: Acrasia Epistémica; Constrangimentos; Racionalidade Limitada

Epistemic akrasia is the theoretical counterpart of a phenomenon that has sparked the curiosity of philosophers with an interest in practical reasoning since Aristotle. However, when compared to its practical counterpart, epistemic akrasia has travelled a much shorter journey in the history of philosophy. Despite the fact that the debate on epistemic akrasia has only started to gain traction in the last few decades, it has not often (at least, not often enough) resorted to the progresses made over millennia on the problem of practical akrasia (with some exceptions). In this essay, I'll start by surveying the main approaches to epistemic akrasia available in the literature, before presenting my proposal: that a theory that explains practical akrasia by appealing to the way in which beliefs are formed under different kinds of constraints can be successfully expanded to explain cases of epistemic akrasia with the same success. I'll finish by showing how this proposal can then be integrated into the program of bounded rationality, thus providing a framework from within which the rationality of cases of epistemic akrasia can be assessed.

Key-words: Epistemic Akrasia; Constraints; Bounded Rationality.

Acknowledgments

For helpful initial discussions, I thank the members of the Epistemology Reading Group that ran on the University of Lisbon in 2020, in particular Bruno Jacinto for organizing the reading group.

For guidance, not only of this thesis, but since the early stages of my B.A., I thank my supervisor, António Zilhão.

And for supporting my education, I thank my parents.

Contents

Introduction

Philosophers have theorized about akratic phenomena at least since Plato's *Protagoras*. Akrasia is usually presented as a combination of a desire, a belief on how to satisfy that desire and a performed action. An example of an akratic situation is:

> Mary wants to get a degree in Mechanical Engineering. She narrows down the options to a couple of universities in Lisbon, the University of Lisbon and the Nova University Lisbon. After careful deliberation, and discussing with friends and family, she comes to the conclusion that, all things considered, she should apply to the Nova University Lisbon. However, when the time comes, she finds herself applying to the University of Lisbon instead, even though she still thinks that Nova was a better option.

There is some sort of mismatch between Mary's best belief on how to satisfy her desire for a degree in Mechanical Engineering and the action she performs. This kind of mismatch is what is characteristic of akrasia. The general schema of akrasia is:

1. *S* desires *K*.
2. *S* believes that, all things considered, the best way to get *K* is to do *A*.
3. *S* does *B*.

Cases of akrasia are, therefore, presented as cases in which an agent's action goes against her best belief on what to do. As such, they are typically seen as cases of conflict, and the akratic agent is, often, characterized as being surprised by her own actions and to struggle to make sense of her own actions (Davidson, 1969: 42). The kind of conflict presented by cases of akrasia is quite puzzling, and, understandably, appealing to philosophers. How can it be that an agent, which has considered every reason that she has to consider, still acts against a belief formed in that way? Some, like Plato himself (*Protagoras* 358b-c), argued for the impossibility of akrasia. Others, like Davidson (1969), have tried to explain its possibility, and evaluate its rationality. The majority of the philosophers after Davidson agree with his point that akrasia is real, and have argued for refined versions of Davidson's explanation or for alternative accounts of the phenomena. For the remainder of this thesis, I shall assume that this Davidsonian way of tackling the issue is the correct one: there are genuine cases of akrasia, and what we, as philosophers, have to do is explain its possibility and evaluate its rational status.

Even though the debate on akrasia is over two thousand years old, it wasn't until forty years ago (Rorty, 1983) that its theoretical counterpart started to receive some attention

by philosophers, and the debate has only really gained traction in the last decade. Akrasia, as discussed by Plato and Davidson, is a phenomenon of the practical realm, i.e., one that involves what agents do. Epistemic akrasia, on the other hand, is a phenomenon of the theoretical realm: it involves only the thoughts of agents. Here is one example of a case of epistemic akrasia:

> Emily is a young student of Philosophy. She is also a catholic. Naturally interested in the question of God's existence, Emily has dedicated most of her studies to this subject. She read medievals (Anselm, Bonaventure, Aquinas, Duns Scotus,...), moderns (Descartes, Leibniz, Hume, Kant,...) and contemporaries (Plantinga, Mackie, Swinburne, van Inwagen,...). After critically reflecting on her readings, she concludes that she does not have good reasons to believe in God's existence. Furthermore, she strongly believes she has good reasons to believe in God's non-existence. Nevertheless, she keeps her belief in God's existence.

The general schema of epistemic akrasia is:

1. *S* believes that, all things considered, she should believe that *p*.
2. *S* believes that *q*,

where *q* expresses a different proposition than *p*. Cases of epistemic akrasia are, therefore, cases of conflict, not between a belief and an action, but between beliefs in the same belief system. Just like akrasia strikes one as either impossible or, at least, puzzling to explain, so does epistemic akrasia. In this thesis I'll survey several accounts of epistemic akrasia. I'll start by considering a sceptic account, according to which there is no distinct phenomenon to be classified as epistemic akrasia, and that what we are classifying as epistemic akrasia are, in fact, cases of a much better known philosophical puzzle (one which has a much older printing history than epistemic akrasia). I will then move on to evaluate several accounts of epistemic akrasia. The goal of this survey of options is not to defend or object to any of them (even though I will do that at points) but rather to inform the presentation of my own account of epistemic akrasia. One way to look at this initial survey and at the argumentative structure of this thesis is this: my own account only gains colours when contrasted with the alternatives in the literature, and its advantages only gain significance if the advantages and disadvantages of the alternatives have been fleshed out in advance.

Before ending this introduction, I'd like to briefly address a question I asked myself several times, specially in the beginning of this project. Why should you care? Isn't this just one of those philosophical brain teasers, with no relation to questions that actually matter? Isn't the debate just appealing to those that like to play this particular puzzling game? Emily's example was specifically designed to, in part, answer these questions, and to appeal to a broader audience. Emily is committed to some kind of fideism, a position in the epistemology of religion which states that belief in God is not to be justified by appealing to reason, but to faith. This is a position that, regardless of its philosophical successes or failures, finds much acceptance outside philosophical circles. Emily is not alone in believing that she has no good reasons to believe in God's existence and, nevertheless, maintain that belief. If this is a common position, and if agents adopting this

position are epistemically akratic, then the debate on epistemic akrasia should be appealing to a broader audience. Understanding what enables one to come to such a position, and, perhaps more importantly, understanding whether it is rational to maintain that position should appeal to those that find this position attractive, and to those that, albeit not attracted by this particular position, are interested in debates in the epistemology of religion. I am sure other examples could be presented to broaden my audience even more. But, for now, this will suffice.

Chapter 1

Surveying the options

1.1 Moore's problem and epistemic Akrasia

On a first encounter, the problem of epistemic akrasia might sound familiar to some philosophers. It may remind you of another problem, one which has motivated much more writing and discussion than the one I want to discuss in this essay. That familiar problem is known as "Moore's problem" or "Moore's paradox".[1] Moore himself presented the problem through a couple of examples:

> Although it may be true both that I went to the pictures last Tuesday and that today I don't believe that I did, it would be 'perfectly absurd' for me to assert the sentence 'I went to the pictures last Tuesday, but I don't believe that I did'. (Moore, 1942: 543)

> I believe he left, but he didn't do it. (Moore, 1944: 175-176)

Moore is trying to motivate a view according to which all utterances of the forms 'p and I don't believe that p' or 'p and I believe that not p' are absurd, not in virtue of the mere utterance of the words, but in virtue of the assertive nature of our utterance of sentences (Moore, 1993: 207). Furthermore, Moore suggests that the absurdity of these assertions manifests itself in the fact that when we utter such sentences, we tend to ask ourselves how it is possible that the assertion of a meaningul conjunctive whose conjuncts can both be true in many occasions be absurd (Zilhão, 2017: 381). Schematically, we can present the problem as follows:

1. It can be true, in a certain moment *t*, both that p and that I don't believe that p (or that I believe that not p).
2. I can assert, in a certain moment *t*, both that p or that I don't believe that p (or that I believe that not p).

[1] The literature favours the use of the phrase "Moore's paradox". I will, instead, use the alternative "Moore's problem". My main reason for doing this is that, as I see it, naming the problem 'Moore's paradox' already involves a commitment to a spectrum of approaches to the problem, namely, those approaches that assume that there is an intuition of paradox in need of explanation. The phrase "Moore's problem" allows me to remain neutral; which is something desirable, as it is not my goal to discuss Moore's problem in any good lenght here, but rather to compare it with epistemic akrasia before quickly returning to proposals of characterization of this latter phenomenon.

3. I cannot assert, in a same moment *t*, that p and that I don't believe that p (or that I believe that not p).

Now that we have a feeling of what's at stake in the discussion of Moore's problem, it is time to ask "Why should we care if Moore's problem is similar to the problem of epistemic akrasia?". There are a couple of reasons to do so. The first is a worry that the problem of epistemic akrasia might colapse into Moore's problem. If the two problems are one and the same, then there is not a distinct phenomenon which this essay seeks to characterize and evaluate: 'problem of epistemic akrasia' would just be another phrase for Moore's problem. I should both rename this essay and turn my attention to the vast literature on Moore's problem. The second reason is that, even if the two problems do not coincide, there may be similarities between the two problems and, therefore, engaging with some literature about Moore's problem might help illuminate the problem of epistemic akrasia. This motivates us into looking into the resemblance between the two problems more thoughtfully.

To start, let's compare the characterization of Moore's problem with my preliminary characterization of epistemic akrasia:

> An agent S is epistemically akratic when S holds the following combination of attitudes: i) S believes that p and ii) S believes that she should not believe that p.

Apparently, there is an obvious difference between the two problems. Moore's problem is about assertions, whereas the problem of epistemic akrasia is about beliefs. But the questions now are "Does this apparent difference amount to a substantial difference? Or is talk of assertions and beliefs interchangeable?". What we need, if we want to defend that there is a difference between the two phenomena, is a precise way of making this distinction.

Owens (2002) suggests one way of doing so. His main idea is that the "requirements of fact" and the "requirements of evidence" can come apart, and that it is this divide that explains the difference between the two problems. Here are his examples, to help make his point clear:

> (1) I believe Jones is innocent but this belief is based on insufficient evidence.
>
> (...)
>
> (2) I believe Jones is innocent but he is guilty. (Owens, 2002: 382)

(1) can be seen as a case of epistemic akrasia: the agent is in a state in which she both believes that p (that Jones is innocent) and that she shouldn't believe that p (because p is based on insufficient evidence). On the other hand, (2) is the sort of claim that Moore discussed: the agent claims both that she believes that p and that p is false. Owens argues that in (2) the intuition of incoherence derives from the simultaneous "committing and evading" nature of the claim; but that no such source of incoherence is found in (1): the epistemic state expressed in (1) is certainly not ideal, but it does not strike us as impossible. You can come to (1) by, despite having received information to the contrary, you still can't shake the belief that Jones is innocent (Owens, 2002: 382); but it is quite harder to imagine how you could come to assert something like (2). In (1), you go against the requirements of evidence and in (2) you go against the requirements of fact. It seems like

the paradoxical feature of (2) is rooted in it being contrary to the requirements of fact. That no such intuition of incoherence is found in (1), shows that going against the requirements of evidence is, somehow, a lighter offence than going against the requirements of fact. And if (2) does commit this heavier offence, while (1) does not, then if (1) is in fact incoherent or paradoxical, it doesn't seem to be because it is equivalent to (2).

The same can (and should) be said about alternative ways of presenting similar states:

> (3) The evidence is sufficient to establish Jones's guilt but I just can't believe that he is guilty.
>
> (...)
>
> (4) Jones is guilty but I don't believe it. (Owens, 2002: 383)

Owens argues that in this case, just like in the first one, (4) goes against stronger requirements than (3). In (4), the agent starts by claiming that Jones is guilty and, "in the same breath", takes it back (by claiming that she doesn't believe it); but in (3), she (merely) claims that were she being reasonable, she would believe in Jones's guilt and that she is not, in this particular instance, being reasonable. It is not an ideal epistemic state, but it is not as bad as the one described by (4). As a saving grace, at least in (3) the agent seems to recognize that something is not quite right about the state she is in. Owens conclusion is that it seems to be possible to believe "in the teeth of the evidence" (Owens, 2002: 384), but it doesn't seem possible that one can believe "in the teeth of the facts". So if epistemic akrasia is impossible, it is not because Moorean states are impossible. That would be the case if Moorean states and epistemically akratic states were equivalent. But the fact that the kind of requirements they fail to satisfy are different seems to show that such an equivalence relation does not obtain.

Owens arguments are useful, not only to dismiss arguments that try to rule out the possibility of epistemic akrasia by establishing an equivalence with Moore's problem, but because they also help us get a grip on what is, or should be, at stake in the discussion of epistemic akrasia. Owens distinction between the "requirements of fact" and the "requirements of evidence" helps to make it clear that a good characterization of epistemic akrasia should appeal to the way in which evidence is formed, received or treated (to the ways in which an agent interacts with evidence), and does not, necessarily, have to say something about truth. All the proposals of characterization of epistemic akrasia that I'll discuss in the next chapters meet this requirement: the proposed characterization is made at the expense of an explanation of how something about evidence works.

1.2 Critical reasoning and falibilism

In this chapter, I'll first present two views that seek to characterize epistemic akrasia by appealing to fallibilist considerations. I'll then criticize both views at once. This will make sense because, even though the proposals differ in some significant aspects, they are vulnerable to the same objections.

1.2.1 Critical reasoning and epistemic akrasia

Borgoni and Luthra (2017) propose to characterize epistemic akrasia at the expense of a fallibility constraint on critical reasoning. Their slogan is: "epistemic akrasia is possible because we are fallible as critical epistemic reasoners" (Borgoni and Luthra, 2017: 886). They start by characterizing critical reasoning as that reasoning in which one goes from a belief that one ought to believe *p* to a belief that *p* (Borgoni and Luthra, 2017: 878). Then, they characterize *epistemic* critical reasoning as that kind of critical reasoning in which "one explicitly takes account of one's epistemic situation" (Borgoni and Luthra, 2017: 878). From here, they present an argument to show that cases of epistemic akrasia are cases of epistemic critical reasoning in which the agent fails to perform the right transition from the epistemic prescription to the belief in what is prescribed. The argument goes as follows:

> 1. Critical epistemic reasoning, attributable to the whole individual, is possible.
>
> 2. The final step in critical epistemic reasoning is a step from a belief that marks acceptance of an all-things-considered explicit epistemic prescription in favor of believing that p to a belief that p, through your appreciating that the epistemic prescription rationally requires the belief.
>
> 3. There is an error constraint on critical reasoning: being guided by an explicit prescription, as one is in critical epistemic reasoning, requires that there is a possible scenario of violating the accepted prescription.
>
> 4. In violating the prescription accepted in critical epistemic reasoning, there is the kind of mismatch that is characteristic of epistemic akrasia. So that kind of mismatch is possible.
>
> 5. In the scenario of violating the accepted prescription, the mismatch is due to the fallibility of a critical reasoner. As a result, it is within the individual's power to avoid the mismatch directly through critical reasoning. In doing so, she would avoid the mismatch as a direct result of appreciating that the epistemic prescription rationally requires the belief.
>
> 6. So epistemic akrasia is possible. (Borgoni and Luthra, 2017: 880-881)

The main point of their argument concerns the link between reasoning and fallibility (premise 3). This is the premise they spend most of the article defending. They start by noting how the premise they make use of is a weaker version of a widely accepted principle, although most often implicitly. That is the principle that all reasoning requires the possibility of error (Borgoni and Luthra, 2017: 882). One can get the intuitive appeal of the idea by considering the following line of thought: one is responsible for one's own reasoning in such a way that one's reasoning is either praiseworthy, blameworthy

or neutral; it only makes sense to say that one's reasoning is praiseworthy or blameworthy if it were possible to have done otherwise; if it is possible to have done otherwise, then it is possible to err; so, if reasoning is susceptible of being evaluated, then it requires the possibility of error; reasoning does seem to be susceptible of being evaluated (as evidenced by our every-day practices); so, reasoning requires the possibility of error. If this is true, than Borgoni and Luthra's claim that critical reasoning requires the possibility of error is also true: it is a particular case of the broader principle.

Another argument presented by Borgoni and Luthra to defend their claim comes from the nature of prescriptions. Their contention is that one does not even count as responding to a prescription, command or imperative if it is not possible to err (Borgoni and Luthra, 2017: 883). Critical epistemic reasoning, they argue, is a way of responding to a prescription, namely, an epistemic prescription (as I said above, they conceive critical epistemic reason as a process of going from a belief in an epistemic ought to the belief in what is prescribed). So if we are capable of engaging in critical epistemic reasoning, then it is possible for us to err in doing so. Once again, their restricted claim is derived from a broader claim, and it is that broader claim that is supported by the arguments presented. A putative objection to this argument contends that this claim about the nature of prescriptions is only true in the case of *practical prescriptions*, i.e., prescriptions on action. That is because Borgoni and Luthra's argument seems to rely on the assumption that we respond *voluntarily* to prescriptions, and that kind of voluntary responses are only found in the practical domain: "since responses in epistemic reasoning to epistemic prescriptions are non-voluntary, that responsiveness to prescriptions does not require the possibility of error" (Borgoni and Luthra, 2017: 883). Their first response to this objection seems rather unsatisfactory. Their point is that because the objection relies on the assumption that responding to prescriptions requires voluntary control, then the objection should show that there are no genuine epistemic prescriptions, rather than showing that we cannot voluntarily respond to them. The reason why this response is unsatisfactory is because this could serve their opponents just the same: showing that there are no genuine epistemic prescriptions rebuts the argument from the nature of prescriptions just as well as showing that these prescriptions do not require the possibility of error. Their second response is better. They cast doubt on the claim that responding to prescriptions, of whichever sort, requires voluntary control. Their point is that we are capable of exercising our voluntary control only over our choices, and we rarely, if ever, choose our choices and intentions (Borgoni and Luthra, 2017: 884). Their conclusion, then, is that our responsiveness to prescriptions is rarely, if ever voluntary and, so, that if prescriptions do require the possibility of error, as Borgoni and Luthra think they do, then it is not because voluntary control is involved.

1.2.2 Higher-order fallibility and epistemic akrasia

Daoust (2019) proposes an account of epistemic akrasia that is, like Borgoni and Luthra's, *fallibility-centred*. His main concern is with defending a solution to rational puzzles involving akratic constraints. I will leave the evaluation of that solution to the second part of this essay. In this section, I will try to reconstruct his characterization of epistemic akrasia from hints in his paper. Since he does not explicitly endorse this characterization, phrases like "Daoust's characterization" should be understood as "my characterization based on Daoust's paper".

Daoust starts by saying that cases of epistemic akrasia are cases of *level-splitting*, that is, cases in which an agent has sufficient reasons to believe that she has sufficient reasons

to believe that *p*, while not having sufficient reasons to believe that *p* - cases in which first-order reasons and higher-order reasons come appart (Daoust, 2019: 288-289). He then explores two explanations of level-splittingness. The first explanation appeals to *incommensurability*.

> **Incommensurability.** Epistemic reasons to believe P and epistemic reasons concerning what one has sufficient reason to believe are incommensurable. In such a case, the balance of epistemic reasons to believe P differs from the balance of reasons for believing that one has sufficient epistemic reason to believe P. (Daoust, 2019: 289)

His reasoning behind this is that if first and higher-order reasons were always commensurable, then reasons for believing *p* would be reasons to believe that one has reasons to believe *p* and vice-versa. However, this is not the case. In certain situations, some reasons might affect my higher-order belief and yet have no effect on my first-order beliefs. Daoust suggests that the following is one such case:

> **Bad Reasoning.** Watson concludes that the killer is Jack the Ripper on the basis of numerous distinctive features X (the type of murder, the type of victim, the crime scene's location, etc.). However, he also has evidence (i) that Holmes thinks that he (Watson) made a mistake in processing the evidence and (ii) that Holmes is almost always reliable. For example, Holmes could suggest that, on that particular occasion, Watson reached a conclusion through incorrect reasoning. (Daoust, 2019: 282)

He suggests that it is plausible to think that in this case Watson could draw the higher-order conclusion that he does not have sufficient reasons to believe that the killer is Jack the Ripper, without leading him to any other conclusion about who the killer is (Daoust, 2019: 289). Incommensurability allows, and explains, this type of cases. However, as Daoust points out, this is not a very convincing argument for level-splittingness. Even if we think that in certain cases higher-order reasons have no effect on first-order beliefs, it is highly implausible to claim that they never do. It is just as likely, if not more likely, to think that Watson could have drawn not only the higher-order conclusion that he does not have sufficient reasons to believe that the killer is Jack the Ripper, but also the corresponding first-order conclusion that it is not the case that the killer is Jack the Ripper.

Daoust then proposes an alternative argument for level-splittingness:

> **Higher-Order Fallibilism**. One can have fallible sufficient reason for believing that one has sufficient reason to believe P. In a case where such a reason is misleading, it is possible that one is rational to conclude that he or she has sufficient epistemic reason to believe P while lacking sufficient reason for the belief that P. (Daoust, 2019: 290)

The rationale for this explanation of level-splittingness is that if higher-order reasons are fallible, then having sufficient reasons to believe that one has sufficient reasons to believe that *p* does not entail that one has sufficient reasons to believe that *p*, which can, in turn, lead to level-splittingness. It is the fallible nature of higher-order reasons that makes room for cases in which an agent's first-order beliefs do not align with her higher-order ones. Cases of level-splittingness, then, can be understood as cases in which the

agent doubts her own higher-order judgements. He then hints that higher-order fallibilism is at the root of cases of epistemic akrasia, by claiming that higher-order fallibilism is the reason why rational puzzles involving epistemic akrasia hold (Daoust, 2019: 291). We can get closer to the way in which Daoust is conceiving epistemic akrasia by considering the argument he provides to deal with a puzzle about a conflict between rationality constraints:

> (1) There can be cases of level-splitting only if agents respond to higher-order fallible reasons.
>
> (2) Relative to the probabilistic representation of reasons, higher-order fallible reasons can be represented as conditional probabilities and higher-order infallible reasons can be represented as unconditional probabilities.
>
> (3) But conditional probabilities can be replaced by unconditional probabilities.
>
> (4) So, relative to the probabilistic representation of reasons, fallible higher-order reasons can be replaced by infallible higher-order reasons, and agents can avoid responding to fallible higher-order reasons.
>
> (C) So, relative to the probabilistic representation of reasons, cases of level-splitting can be avoided. (Daoust, 2019: 291)

This argument, coupled with *higher-order fallibilism*, hints at the following characterization of epistemic akrasia:

> **Higher-order fallibility and epistemic akrasia:** cases of epistemic akrasia are cases made possible by the fallibility of higher-order reasons. An agent comes to hold an epistemically akratic combination of beliefs when she fails to appeal to the relevant higher-order infallible reasons.

Much like Borgoni and Luthra's, Dauost's characterization depicts epistemic akrasia as a type of fallibility-centred mental failure. This is why, in the beginning of the chapter, I said that I would criticize both views simultaneously: my objections are directed at all accounts of epistemic akrasia that characterize it as a failure.

1.2.3 Objections to fallibility-centred proposals

My main objection to fallibility-centred characterizations of epistemic akrasia comes from an analogy with cases of practical akrasia. The point is that fallibility-centred characterizations err in conceiving cases of akrasia as failures, since there are genuine cases of akrasia which don't seem to constitute failures. To make my point clearer, lets start with this example:

> Let me invoke a very simple example to illustrate my hypothesis. Suppose that you are taking part in a conference in a foreign country and, at lunch time, which is served in a self-service, and after having eaten the main course, you're in the line for dessert. There are two possible choices, but you're only familiar with one of them: a piece of fruit which is equally common in your place of origin. Now imagine that that kind of fruit doesn't spark great enthusiasm in you, even though you don't find it unpleasant either. The other possible choice is a sweet. You are not under any dietetic constraints, you

> like sweets and you want one for dessert. But this one is unfamiliar to you. You can recognize some of the ingredients as being to your taste, but others you don't know at all. The sweet looks, however, vaguely appetizing. Now suppose that you initiate an explicit deliberation process in order to make your choice. While you wait in line for your turn to pick a dessert, you decide to pick the sweet. After all, you think, if you later discover that it is not to your taste, you'll end up with no dessert to finish your lunch and will have lost a couple of euros, neither of which is a particularly tragic event. But, if it is to your taste, you will have finished your meal in a specially pleasant way, which, at the moment, is what counts most. However, when your turn to reach for the desserts comes, you pick up one of the exposed pieces of fruit. Even though you realized what happened, your action still surprises you. (Zilhão, 2010a: 319-320)

This is a case of practical akrasia in which the agent acts against her own best judgement. She judges that the best course of action is to pick the sweet, but ends up picking a piece of fruit. However, and this is what is crucial to my argument, this should not be seen as a failure. Zilhão uses this example to illustrate how, in many cases, the akratic action is objectively better action (Zilhão, 2010a: 319). She acted against her own best judgement, but that doesn't mean that she didn't perform the objectively better action that she could perform. Zilhão then goes on to explain how psychological mechanisms, namely, a *recognition heuristic*, triggered the action contrary to the agent's explicit judgement. The point to retain is: even though cases of practical akrasia are always cases in which one acts against one's own best judgement, they are not necessarily cases of failures. Being practically akratic could, in certain situations, prove essential to one's survival. So it seems that accounts of practical akrasia that characterize it at the expense of a failure leave out cases like this one, which, in turn, means that they are bad characterizations of practical akrasia because they do not account for all cases of practical akrasia. Assuming that cases of epistemic akrasia are analogue to cases of practical akrasia (as I'll argue in chapter 4), then it would be good if we avoided accounts of epistemic akrasia that characterize it as a failure.

Another cause for concern about these fallibility-centred accounts of epistemic akrasia is that they do not shed much (if any) light on what is going on in the subject when she finds herself on an akratic situation. Borgoni and Luthra's argument for the possibility of epistemic akrasia states that it is possible for reasoners like us to be epistemically akratic because we can make the sort of mistakes that lead to an akratic combination of attitudes. However, the more interesting question, which other proposals will try to adress, is why can we make such mistakes. What is it about the way our mind works that allows us to split our attitudes in such a way as to arrive at an akratic situation. This is not a devastating argument for fallibilists, as their proposals could be enriched with an explanation for that fallibility; but it is worth noting that fallibility by itself does little to enlighten the question of how we can be epistemically akratic.

1.3 Fragmentation

The most influential proposals for characterizing epistemic akrasia have been made along the lines of what we might call "fragmentation views". Fragmentation theorists hold that certain epistemic phenomena are best explained by a theory of human mind that depicts it as being composed of several fragments.[2] The theories I'm interested in are those that explain epistemic akrasia by appealing to a fragmented view of cognition or mind (regardless of their treatment of other epistemic phenomena). I will first present the most influential version of this characterization of epistemic akrasia, then consider a proposed refinement and end with some criticism of these views.

1.3.1 Greco's view

Greco (2014) characterizes epistemic akrasia at the expense of a multiplicity of belief formation systems. Greco suggests that the mind is fragmented in the sense that we have linguistic and non-linguistic systems dedicated to the formation of beliefs.

> For example, imagine a gambler who is disposed to bet on red if the last 10 spins of the roulette wheel have come up black, unless she says to herself something like the following: ''The gambler's fallacy is a fallacy, and the fact that black has come up many times recently doesn't mean that red is any more likely to come up the next time.'' Suppose that when she does remind herself of these facts, she bets in line with the true probabilities, rather than the ones that would be suggested by the gambler's fallacy. In such a case, we might distinguish between the dispositions produced by her non-linguistic belief system—call these ''beliefs$_n$''—and the dispositions produced by her linguistically infused belief system—call these ''beliefs$_l$''. We might go on to say that she believes$_n$ that red is more likely to come up after a string of blacks, but at the same time she believes$_l$ that this is not the case, and that the chances of red coming up on any given spin are independent of the results of the previous spins. (Greco, 2014: 209)

Greco admits that it isn't obvious that we should describe the gambler's case as a case of epistemic akrasia. That is because it is not clear that we should ascribe the gambler both the straightforward belief that red is more likely to come up after a run of blacks and the straightforward belief that she shouldn't believe this (because she is aware that the gambler's fallacy is, indeed, a fallacy). He argues that the distinction between beliefs formed by distinct belief formation systems allows us to meaningfully make this description.

> However, once we have made a distinction between beliefs$_n$ and beliefs$_l$, we should allow that in some cases, an agent might be disposed by her linguistically infused system to act as if P, but this disposition might be masked by the operation of her non-linguistic belief system, which we may suppose even more powerfully disposes her to act as if P. (Greco, 2014: 209)

The possibility of epistemic akrasia is assured by the possibility of conflict between different belief formation systems. It is because the diferent systems can produce conflicting dispositions that human agents can be described as having conflicting beliefs like

[2] For examples of fragmentation views applied to other phenomena, see: Lewis (1982), Egan (2008) and Stalnaker (1999).

the ones required to be in a state of epistemic akrasia. Furthermore, Greco stresses that, despite talking about linguistic and non-linguistic belief formation systems, nothing serious rests on this. Talk of linguistic and non-linguistic systems is merely a way to simplify his discourse. What really matters is that we are able to identify at least two different belief formation systems and that these belief formation systems can produce conflicting verdicts.

Greco argues that his explanation of epistemic akrasia is not metaepistemologically neutral (Greco, 2014: 210). That is, his account of epistemic akrasia depends on a certain view of the metaphysics of epistemology, in particular, a certain view of the content of epistemological beliefs (that is, beliefs about epistemological matters). His contention is that his view of epistemic akrasia suggests a metaepistemology that is similar to what in metaethics is called "expressivism". Metaethical expressivists argue that moral beliefs or moral statements are not to be analyzed as carriers of cognitive content, but rather as expressions of non-cognitive states.[3] Likewise, metaepistemological expressivists hold that epistemological beliefs or statements are not vehicles of cognitive content, but expressions of something else.

> That is, rather than understanding epistemological beliefs as having a distinctive sort of content, we could try to understand them in some other way. In particular, we might understand them as distinctive in terms of the psychological mechanisms that produce and maintain them, rather than as distinctive in terms of their subject matter. The belief that my evidence supports the claim that I have hands, then, could be understood as having the same content or subject matter as the belief that I have hands; it would be distinctive only in being produced or maintained by a distinctive sort of psychological system (e.g., a linguistically infused system). (Greco, 2014: 210-211)

The main upshot of the expressivist's position (and the main point of controversy with her rivals) is that the subject matter of a belief about p and the subject matter of a belief about the belief about p are one and the same. That is, epistemological beliefs are not about a distinct subject matter (such as the topic about relations of inferential support, or the one about the probatory powers of testimony), but about the very same things as their first-order counterparts. Take two beliefs: B_1) I believe that it is raining; B_2) I believe that my evidence supports my belief that it is raining. One natural way to describe my epistemic state is as one in which I have two distinct beliefs about two distinct subject matters (or topics or questions). B_1 is about the weather (or, if you want to treat questions more finely, about whether or not it is raining) whereas B_2 is about what my evidence supports (or, again, about whether or not my evidence supports my belief that it is raining). But this description requires the attribution of a cognitive content to B_2, which expressivists like Greco would deny. For them, both B_1 and B_2 are about the weather (or about whether or not it is raining). B_2 is just another of believing the content of B_1.

In the previous example, I presented expressivism as a rival for a natural description of epistemic states with beliefs of different orders. The purpose of such presentation was to suggest that expressivists have to tackle several problems in order to make their position worthy of consideration. Greco is aware of that, and tries to do deal with two problems

[3] Emotivists, like Ayer (1971), are expressivists who believe that the non-cognitive states expressed by moral beliefs or statements are feelings or emotions. For an overview of non-cognitive approaches to morality, see (Miller, 2010).

in his paper: *the negation problem* and *the possibility of justified false belief problem*. I'll follow Greco's presentation and start with the negation problem.

> **The Negation Problem:** i) what kind of belief is a belief such as "I ought not to believe that p"?; ii) how do we preserve (if at all) the distinction between "I ought not to believe that p" and "It is not the case that I ought to believe that p"?

The difficulty with the first thing that the problem asks the expressivist to explain is that, since she reads "I ought to believe that p" as a kind of believing that p, it would seem natural to read "I ought not to believe that p" as another kind of believing that p. But that is weird: the belief is about an epistemic ought not to believe p; why would we think that it is a disguised way of believing p? But then again, how is the expressivist going to explain this without appealing to a distinct (epistemological) subject matter? The second thing that the poser of the problem requires an explanation for is a further worry of the same kind of the first one. The problem is how to account for intuitively different epistemic evaluations of the same propositional content.

Greco's answer to the first problem consists in showing how we can preserve the contrast between intuitively different beliefs without appealing to a distinct subject matter. One way of doing so is by appealing to the notion of *disagreement* (Gibbard, 2003). The idea is that those intuitively different beliefs have different *disagreement profiles*, that is, they are incompatible with different mental states.

> The belief that one's evidence supports the belief that P (...) disagrees both with believing P, and with being agnostic concerning P. The belief that it is not the case that one's evidence supports the belief that P (...) will have a different disagreement profile; it will disagree with believing that P, but will fail to disagree with being agnostic concerning P. (Greco, 2014: 211)

Greco's (and Gibbard's) contention is that disagreement profiles allow us to maintain the distinctions between "I ought to believe that p", "I ought not to believe that p" and "It is not the case that I ought to believe that p". This addresses the second part of the first problem. But, Greco tells us, it also helps with the first worry posed by this problem (how to characterize beliefs such as "I ought not to believe that p"). If we characterize beliefs by appealing to their disagreement profile, instead of their intentional object (of what they are about), we can avoid the kind of paradoxical claims that motivated the first problem for expressivism.

Now the second problem:

> **The Possibility of Justified Belief Problem:** how is it possible for an agent to have a justified false belief if both her false belief and the belief that she ought to have such belief are about the same subject matter?

Here is how Greco presents it:

> Suppose you believe (non-akratically) that your car is parked in your garage, where you left it. As it turns out, however, I have recently stolen your car and sold it to my neighborhood chop-shop. Now, consider what I should think about the following two beliefs of yours:

• Your car is parked in your garage.

• You ought to believe that your car is parked in your garage.

> It seems that I should think that the first belief of yours is false, while the second belief is true—I should take you to have a justified false belief. But (and this is the puzzle) if I am to coherently think this, then it seems as if your two beliefs must concern different subject matters—they can't both just be beliefs about where your car is parked, because then there would be no way for one of them to be correct while the other is incorrect. (...) More generally, how can we account for the possibility of justified false belief without positing a distinctive epistemological subject matter? (Greco, 2014: 211-212)

The challenge, then, is to show how an expressivist can account for the possibility of justified false beliefs without committing herself to a distinct epistemological subject matter. Greco proposes that, in the spirit of expressivism, he will answer this problem by treating normativity indirectly, by explaining what is involved in making normative judgements; that is, instead of explaining how are justified false beliefs possible, he explains what is involved in *thinking* that someone has a justified false belief (Greco, 2014: 212). Greco notes that typical cases of attribution of justified false beliefs are third-personal, and, so, that what he needs to explain is what is involved in thinking that *someone else* has a justified false belief (i.e., he needs to explain what is involved in believing "She ought to believe that p but $\neg p$").

With that in mind, Greco suggests that an expressivist's account of such beliefs should be akin to her account of beliefs of the form "I ought to believe that p": just as an expressivist explains this latter belief by appealing to the agent's beliefs about p formed by linguistic systems, so the expressivist explains the former belief, this time by appealing to a set of linguistic beliefs possessed by the agent under evaluation.

> (...) it's plausible that just as an expressivist account of what it is for me to believe that I now ought to believe that P will turn on my current beliefs$_{\text{l}}$ about whether P, an expressivist account of what it is for me to believe that someone else ought to believe that P will turn on something like my conditional beliefs$_{\text{l}}$ about P, given only information available to the relevant person. That is, I might think that P, while thinking that some body of evidence E (not my evidence) supports believing that ¬ P. This might involve believing$_{\text{l}}$ that P, while having a kind of conditional belief$_{\text{l}}$ in ¬ P given E. (Greco, 2014: 212)

Talk of conditional beliefs is a more familiar way of talking of Gibbard's (2003) *contingency plans*. Gibbard's contingency plans for action and belief are plans to perform certain actions or to have certain beliefs if one were to find oneself in the relevant conditions. Saying that I have a belief in "She ought to believe that p" conditional on some evidence E possessed by her amounts to saying that, were I in the same conditions as she is (having the same epistemic capabilities and the same body of evidence E), I would believe that p.

1.3.2 Kearl's refinement

Kearl (2020) proposes a refinement of Greco's view. He says that the main idea, explaining epistemic akrasia at the expense of fragmentation, is correct but that the way in

which Greco himself developed it is not, in the sense that there are cases of epistemic akrasia which Greco's theory fails to classify as such. These are cases of what Kearl calls "higher-order epistemic akrasia". Here is his example:

> Smith believes that it is rational to believe that P. Jones, who is an expert on the rationality of higher-order beliefs, tells Smith that it is irrational to believe that it is rational to believe that P. On the basis of Jones's testimony, Smith believes that it is irrational to believe that it is rational to believe that P. (Kearl, 2020: 2504-2505)

Cases like this seem to be of the form that characterizes epistemic akrasia. If, as Kearl does, we take Q = "It is rational to believe that P", then Smith holds the following combination of beliefs: i) he believes that Q and ii) he believes that he shouldn't believe that Q. But Greco characterized epistemic akrasia at the expense of a conflict between different kinds of belief formation systems (focusing on the distinction between linguistic and non-linguistic systems). And the problem with higher-order akrasia is that, since both beliefs are about what it is rational to believe, both seem to be formed by the same belief formation system, the linguistic system. So, Smith's combination of attitudes can be specified as: i) he believe$_l$ that Q and ii) he believe$_l$ that he shouldn't believe that Q. Kearl's case is one in which the beliefs seem to be of the correct form, but fail to be formed by distinct systems. So, Kearl says, according to Greco it should not be classified as a case of epistemic akrasia. But, intuitively it seems like a genuine case of epistemic akrasia.

> Higher-Order Akrasia presents a case of epistemic akrasia where conflicting beliefs are formed by one and the same belief-formation system, and so there is no conflict between belief formation systems. If Higher-Order Akrasia genuinely presents a possible case of epistemic akrasia, the Fragmentation Analysis is incorrect. (Kearl, 2020: 2505)

The above quote touches on an important part of Kearl's argument which I have not yet mentioned. Kearl reads Greco's paper as a defense of an analysis or definition of epistemic akrasia. That is, he takes Greco to put forward (and argue for) a set of necessary and suficient conditions which a genuine case of epistemic akrasia satisfies. In that sense, finding a genuine case of epistemic akrasia which does not satisfy all of Greco's conditions is enough to prove him wrong. Kearl presents Greco's view under the label "Fragmentation Analysis" (Kearl, 2020: 2503).

> **Fragmentation Analysis:** It is possible for one to be epistemically akratic when one has multiple belief-formation systems, and it is irrational for one to be epistemically akratic because it is a way of accepting inconsistent plans for belief.

Kearl's contention, then, depends on defending the genuineness of Smith's case. He does so by responding to three objections: i) the objection of psychological implausibility; ii) the objection of mere inconsistency; iii) the objection from the lack of substantial functional differences. Along the way, he proposes modifications to the Fragmentation Analysis that would allow it to accommodate cases of higher-order akrasia.

The objection of psychological implausibility can be presented as follows:

> **Objection of psychological implausibility:** Smith is exceptional. Typically, humans don't have third-order beliefs about their second-order beliefs, and it is implausible that they can have such beliefs.

Kearl provides two answers to this objection. The first one claims that Smith isn't psychological implausible (or, at least, not as implausible as the objector supposes). Kearl aks us to imagine a line of epistemologists with n members. Each epistemologist evaluates the rationality of the belief of the previous element of the line, with the first epistemologist of the line believing p. What we are faced with is a sequence of beliefs of increasingly higher-order. Now suppose that, instead of a line of epistemologists, we have a single epistemologist evaluating her own beliefs. Is there, Kearl asks, any relevant difference that makes us be ok with the first example, but classify the second as exceptional? What this response does is take the explanatory burden away from the defender of higher-order akrasia and give it back to the objector.

His second response starts by conceding that there might be something exceptional with Smith's case (because, unlike us, he can entertain third-order beliefs). Be that as it may, Smith is not outrageously exceptional. If our brains were just a little bit bigger, we would be able to entertain third-order beliefs. So, if our brains were just a little bit bigger, Smith's case would be a genuine counter-example to the Fragmentation Analysis. But, Kearl says, what epistemic akrasia is does not depend on contingent facts about the size of our brains. Therefore, Smith's case being a counter-example to the Fragmentation Analysis does not depend on such facts. Therefore, Smith's case is a genuine counter-example to the Fragmentation Analysis.

Kearl also adds (Kearl, 2020: 2507 and note 9) that denying the possibility of higher-order akrasia based on contingent facts about the size of our brains excludes the possibility of more complicated rational agents having irrational states more complicated. Considerations like this could motivate a revision to the objection. There are two kinds of epistemic akrasia: i) epistemic akrasia in agents like us and ii) epistemic akrasia in agents unlike us. The Fragmentation Analysis is an analysis of i), but not of ii). To which Kearl, once again, provides two answers. The first is that the objection is *ad hoc*. There seem to be no independent reasons to make such a distinction. The second is that such a distinction fails to appreciate the continuity between i) and ii). Which is to say that making the nature of epistemic akrasia depend on the size of our brains is a failure to accommodate more complex cases of irrationality.

The objection of mere inconsistency can be presented as follows:

> **Objection of mere inconsistency:** Smith has inconsistent beliefs but he is not epistemically akratic. For him to be epistemically akratic, different systems must come into conflict.

This objection tries to maintain a hold on the idea that, in order for a phenomenon to be a case of epistemic akrasia, there must be a conflict between different belief formation systems. The lack of such a conflict would render Smith's case as one of mere inconsistency. The strength of this objection varies in accordance with our degree of commitment to the idea that epistemic akrasia must involve such a conflict.

Kearl's general response is that, even if Smith's beliefs are produced by the same system, they are, nevertheless, different ways of believing $p/\neg p$ and that classifying Smith's case as one of mere inconsistency fails to capture the intuition according to which Smith

believes $p/\neg p$ in different ways. In order to accommodate for these different ways of believing, Kearl proposes two revisions to the Fragmentation Analysis. The first one is more conservative, while the second one is somewhat more radical (when compared to Greco's original proposal).

The conservative revision is motivated by the thought that if contemporary psychology and cognitive science are guides to the structures in our brains, then there are many linguistic and non-linguistic systems (Kearl, 2020: 2508). Instead of denoting individual belief formation systems, 'linguistic system' and 'non-linguistic system' would denote *sets* of belief formation systems, with each member of the set (or subsystem) satisfying certain core functional roles (which characterize the set). The possibility of epistemic akrasia would be assured by the possibility of conflict between the various subsystems.

The radical revision is motivated by the thought that it in order to explain the possibility and determine the rational status of epistemic akrasia, it is not enough to look inside our brains: we must add in some external factors, crucially *contexts* and *goals*. By making the contexts and goals to which beliefs are indexed a crucial component of the explanation of the possibility and irrationality of epistemic akrasia, we end up with an account of epistemic akrasia that does not bottom up in an account of the structures in our brains (Kearl, 2020: 2508).

Kearl doesn't argue for the superiority of one revision over the other. Instead, he argues that both revisions suffer from the same flaw: they both seem to be permissive in the way they individuate "ways of believing". If one has functionalist dispositions, one can object to this proliferation of ways of believing based on the inexistence of genuine functional differences between the multiple ways of believing. This leads him to consider the third objection, the objection from the lack of substantive functional differences. Start from a broadly functionalist account of belief. What makes a certain attitude an attitude of belief is determined by its role in the cognitive system in which it is embedded. The sort of roles that tend to characterize beliefs include being in certain causal relations with sensory inputs and behavioural outputs. We can now present the objection:

> **Objection from functional differences:** From a functionalist standpoint, higher-order akrasia is, *prima facie*, impossible for there is nothing to functionally distinguish second-order beliefs about what it is rational to believe from third-order beliefs about what it is rational to believe. And if there is nothing that functionally distinguishes them, then they are not distinct at all. Therefore, Smith's case is not a case of epistemic akrasia, but a case of plain inconsistency.

Kearl's response is that arbitrarily complex ways of believing are not only consistent with functionalism, but required by functionalists to play certain explanatory roles. Explanatory roles are not merely superficial differences, so, Kearl argues, functionalists have to accept higher-order beliefs (thus, accepting the revisions he proposed to the Fragmentation Analysis).

Kearl's example concerns Game Theory. In Game Theory, when applying backwards induction arguments, we rely on assumptions of common knowledge between the players. In order to satisfy such assumptions, each player must know an infinite number of propositions, each more complex than the other. For example, in a centipede game of two players, the players take turns choosing either two coins, thereby ending the game, or one coin, thereby giving the other player a chance to make the same choice (Kearl, 2020: 2511). In such a game, each player must know: that they are both rational; that they both

know that they are both rational; that they both know that they both know that they are both rational; and so on *ad infinitum*.

Game theorists tell us that the Unique Rational Strategy (*URS*) in a centipede game of two players is to take two coins in first play, thereby ending the game. But empirical studies show that players almost never play in accordance with the URS. An attractive explanation for that discrepancy is that, in real situations, players rarely (if ever) satisfy the common knowledge assumption. In a finite game, the common knowledge assumption is excessive. Even so, in any finite game, there is a finite set of knowledge states K such that, if player S satisfies K, then he is rationally required to take two coins. The strategy recommended by game theorists is sensitive to the player's knowledge state (or ignorance; depending on whether you are, or not, a "glass half-full" epistemologist). Kearl claims that this, surely, is a legitimate functional difference: if functionalists did not make a distinction between the players higher-order beliefs, then they could not account for the different strategic recommendations of the game theorist. The mistake of the objection from functional differences, Kearl concludes, was that it did not realize that, whether or not we are functionalists, we need arbitrarily higher-order beliefs for non-superficial theoretical ends: if we deny the possibility of arbitrarily higher-order beliefs, we must deny the legitimacy of the application of backwards inductive reasoning.

1.3.3 Fragmentation and disguised inconsistency

One of the most striking features of both Greco's and Kearl's proposals is their adoption of an expressivist meta-epistemology. The main upshot of this meta-epistemology is that there is no difference in subject matter between a belief in *p* and a belief about a belief in *p*. The belief about a belief in *p* is seen as another mode of expression of the same belief in *p.* This blocks a typical move in the literature concerning the rationality of epistemic akrasia: to claim that what one needs in order to properly assess cases of epistemic akrasia is a principle that governs the relation between first and higher-order beliefs. By stating that both beliefs are at the same level (because they are about the same thing), the expressivist's evaluation of cases of epistemic akrasia cannot appeal to considerations about coherence between levels.

As we are faced with two expressions that differ merely in their etiology, and not in content, the expressivist claims that it is much more attractive to see these cases as cases of irrationality. Greco claims that since the expressivist takes the beliefs that form an akratic combination of attitudes to be about the same subject matter, then she must take them to be *inconsistent* beliefs about the one subject matter, rather than consistent, but somewhat conflicting, beliefs about different subject matters (Greco, 2014: 215). Since it is widely accepted (even if not uncontested) that logical consistency is an ideal of epistemic rationality, once one adopts the expressivist framework, it is natural to draw the conclusion that cases of epistemic akrasia are cases of epistemic irrationality because they are nothing more than cases of logical inconsistency in disguise.

Greco (2014: 215) does concede that the fragmentation view of epistemic akrasia does not, by itself, settle the question concerning the rationality of epistemic akrasia. But he does claim that it *changes* the discussion. His point is that the appeal to certain epistemic principles might render cases of epistemic akrasia as rational or irrational; but those involved in the discussion can no longer appeal to a distinction in content and have to make their claims stick in a framework in which the two states involved are only to be distinguished by their etiology. The fact that Greco places so much emphasis on his claim that

the main upshot of his proposal is that it changes the discussion of the rationality of epistemic akrasia, and the fact that such change, if it is real, is anchored on his expressivist framework, justifies a thorough assessment of his meta-epistemological expressivism. In the next section, my criticisms of Greco will, therefore, be focused on his defence of expressivism.

1.3.4 Criticism of fragmentation views

Now that I have presented the main fragmentation views, it is time to discuss some reasons to reject these views. I'll start by criticizing Kearl's, before moving on to Greco's. This is mainly because, as I will more firmly argue below, Kearl's proposal is logically stronger than Greco's.

The first two points of criticism are directed at Kearl's interpretation of Greco's paper. First, Kearl reads Greco as putting forward an analysis or definition of epistemic akrasia. However, Greco never refers to his work as a proposal of such an analysis. Greco carefully uses terms such as "account" or "characterization". In that sense, Kearl's higher-order objection to Greco's proposal is a bit misdirected. Since Greco is not defending a definition of epistemic akrasia, Kearl's objection loses some of its strenght. It should not be seen as an example that refutes the theory, but rather as one that shows that the initial characterization has room for improvement. Second, Kearl's objection rests on the assumption that, for Greco, it is decisive that the conflict arises between a linguistic system and a non-linguistic system. But Greco says that the only reason he talks about this distinction is for simplicity's sake. He claims that not much rests on these labels and that the possibility of epistemic akrasia is assured so long as we can identify any two belief formation systems.

> Not much rests on the two systems being distinct linguistic and non-linguistic belief systems. As long as we can identify some distinct psychological subsystems, each of which has a right to be called a belief system, then we can understand cases of epistemic akrasia as cases in which the outputs of these distinct subsystems conflict. (Greco, 2014: 210)

This makes Kearl's objection even weaker. Not only is he objecting to a definition that was not put forward, he is objecting to a theory based on a distinction that was originally classified as non-important. Greco's original proposal, with the proviso that a strict distinction between a linguistic system and a non-linguistic system is not essential to it, probably had all the resources it needed to deal with cases of higher-order akrasia. Kearl simply needed to show how Greco had laid the foundations of a theory that could be expanded into a better one, and shouldn't have tried to refute what was a perfectly good starting point.

Now I'll present my main objection to Kearl's account. The objection contends that Kearl's approach to the second and third objections (the objection of mere inconsistency and the objection from functional differences) is misguided. Kearl tries to force the reader to accept the claim that believing "It is rational to believe that p" and believing "It is rational to believe that it is rational to believe that p" are different ways to believe that p. But he fails to do so. The revisions he proposes may go in the right direction, but these do not by themselves solve the problem. Kearl would then need to argue that different subsystems *necessarily* or that different combinations of subsystems and contexts *necessarily* produce different ways of believing the same content *or* provide a criterion (or criteria, should one not suffice) to distinguish cases in which different ways of believing

are produced from cases in which that doesn't happen. For it could be the case that, for example, different subsystems produce the same ways of believing a content. Nothing in Kearl's account prevents it. From the supposition that there are as many subsystems/combinations of systems-contexts as one might possibly need for one's theoretical purposes, it does not follow that those subsystems/combinations always produce the desired outcome - that would be too easy and suspiciously *ad hoc*.

Furthermore, instead of dealing with this problem, Kearl goes on to deal with a problem from a lack of functional differences. But, once again, this seems misdirected. Kearl's point that even the functionalist must accept arbitrarily higher-order beliefs creates bigger problems for himself than for the functionalist! The real problem is how is the expressivist going to account for the differences between those arbitrarily higher-order beliefs. The question Kearl should be asking himself is "How am I going to distinguish between these arbitrarily higher-order beliefs if they all have the same content?". Greco appealed to the notion of disagreement profile to try to deal with what he called the negation problem. Kearl took this problem to a much more complicated level and did absolutely nothing to try to solve it. Even if Kearl's response to the objection from functional differences is satisfactory, the more serious problems are left unattended. Kearl's reader is left with the feeling that Kearl sabotaged Greco: instead of taking his original proposal and improve it, he created even bigger problems for it.

Now, three objections to Greco's account. Both objections are directed at his responses to the problems of expressivism. So my first objection is an objection to Greco's response to the problem of negation and my second objection is as objection to his response to the problem of the possibility of justified false beliefs.

The objection concerning the problem of negation can be split into two: one epistemological and one metaphysical. The epistemological concern can be expressed as follows: how do I determine the disagreement profile of a belief without appealing to its content? Greco says that the disagreement profile of a belief consists of those mental states with which the belief is somehow incompatible. But what is it for a belief to be incompatible with other mental states? Greco does not provide a comprehensive characterization of incompatibility, but he hints that it should be analysed at the expense of notions such as inconsistency or incoherence. But inconsistency and incoherence relations are relations that obtain between contents. These are usually understood as logical or probabilistic relations. If disagreement profiles are to be determined by looking into these sort of relations, then if there is no difference at level of content (as the expressivist claims), why should I expect a difference in the disagreement profiles? If the contents are the same, then so are disagreement profiles. The metaphysical concern is very similar, but on a more fundamental level. The question is: what grounds the disagreement profile? It is not merely a question of how we can determine the disagreement profile, but also a question about the source of that profile. Where does this (according to Greco) essential property of a belief come from? The natural response would be to say that a belief's disagreement profile is grounded by its content. But then it is difficult to see how the distinction between content and disagreement profile could do the work the expressivist wants it to do. For it seems that only differences in content could ground differences in disagreement profiles. If the content is the same, as the expressivist claims, then so should the disagreement profile be! So the appeal to disagreement profiles does not enable us to save the distinctions we wanted to preserve. In short, Greco's response does not seem to solve the problem, but just push it to another level.

The first objection to his response to the possibility of justified belief problem is a re-

quest for clarity. It is not clear from Greco's response whether there is an independent set of rules for the evaluation of the rationality of cases of justified false beliefs (or epistemic states in general). That is, it is not clear whether there is a more general framework within which such evaluations are to be made. Greco says that the evaluation of a case of justified false belief is to be made by appeal to the notion of a contingency plan, a plan of what I would do if I found myself in the relevant situation. But he does not say how is this contingency plan to be formed. If anything goes, then not only cases of justified false beliefs but, possibly, also cases of unjustified false beliefs could be classified as rational. And if this "anything goes" attitude is not even diachronically consistent, then there could be two very similar cases of justified false beliefs which differ in their rational classification. The notion of a contingency plan, on its own, is not capable of saving the day; it must be supplemented with a story about how evaluations of rationality come about.

The second objection concerning the problem of the possibility of justified belief asks Greco why is the evaluation of a case of justified false belief and a case of epistemic akrasia different. Greco argues that epistemic akrasia is irrational because, since the two beliefs are about the same thing, it is akin to holding inconsistent beliefs. The question, then, is why are cases of epistemic akrasia classified as irrational, whereas cases of justified false beliefs are classified as rational? In both cases, we face situations in which the two beliefs, albeit with different origins and expressive powers, are about the same thing. The difficulty in explaining how can one of the beliefs be true while the other is false is present in both cases. If in the case of a justified false belief I can classify the agent's epistemic state as rational, what stops me from doing the same in the case of the akratic agent? Can't I interpret the akratic agent's second-order belief by appealing to the notion of a contingency plan and say something like "If I where in that situation (the situation the akratic agent is in), I would also believe that I should believe that p, even though I believe that $\neg p$; therefore, I should classify his state as rational"? The point is not to force Greco to admit that epistemic akrasia is rational. It is not, as well, to force him to give the same rational evaluation to both cases. It is rather to force him to provide a better explanation of why the evaluations differ. Granted, the objection does not defeat Greco, nor was it supposed to. It is certainly weaker than the objection directed at the problem of negation. But, at least, it shows that there is some explanatory work to be done - and even if this is not a reason to outright reject Greco's proposal, it is a motivation to consider alternatives.

1.4 Misleading evidence

Some authors have suggested that the proper way to characterize and assess cases of epistemic akrasia, is by looking at the nature of the evidence involved. The main claim that is common to the various proposals of this sort is that evidence might be misleading, and that it is this misleading nature of evidence that leads to cases of epistemic akrasia. What these authors disagree about is on whether such a misleading nature renders cases of epistemic akrasia as epistemically rational or irrational. Those that defend the rationality of epistemic akrasia rely on the misleading nature of evidence to excuse the agent; those that defend the irrationality of epistemic akrasia will point out the agent's inability to recognize the misleading nature of the evidence as an epistemic failure. In this chapter, I'll present both sorts of views and evaluate them.

1.4.1 Rationally misleading evidence

Williamson (2011 and 2014) suggests that, in certain cases, one may know a proposition even if one's evidence almost guarantees that one does not know *p*. In (at least some of) those cases, the agent can be adequately described as believing *p* (following the traditional definition of 'knowledge', if one knows *p*, then one believes that *p*) and believing that one should not believe *p* (because the evidence "points in the opposite direction" of *p*), which constitutes a case of epistemic akrasia. So, Williamson is committed to the possibility of cases of epistemic akrasia. Cases of long deductions (Horowitz, 2014: 720) are simple cases of the kind Williamson has in mind. Imagine you're going through a deduction with thousands of premises. If you made no mistakes, and get to the right conclusion, then you know that proposition. However, you are aware of the limitations of your computational skills, and believe that it is almost certain that you made a mistake. In this case, you know the deduced proposition, but your evidence supports, with probability close to 1, that you do not know it. Your evidence is misleading, but it is rationally misleading because to have high confidence in the proposition that one made no mistake in such a long deduction would be to assume that one is infallible.

Lasonen-Aarnio (2014) argues that one can be rationally misled by one's own evidence into an epistemically akratic situation because, in some cases in which higher-order evidence has a defeating effect on object-level beliefs, the norms for rational epistemic conduct may suggest conflicting verdicts and there is no good way to resolve this conflict. The point is that different levels of epistemic engagement have different sets of epistemic norms: first-order epistemic agency is subject to first-order epistemic norms, higher-order epistemic agency is subject to higher-order epistemic norms. Suppose that, for each epistemic level $\mathfrak{L}$, there is an epistemic norm that says "Believe in accordance with $\mathfrak{L} - evidence$". In cases in which higher-order evidence has a defeating effect on lower-level beliefs, the agent is presented with conflicting norms of belief: 1) "Believe in accordance with your lower-level evidence" and 2) "Believe in accordance with your higher-level evidence". As we know that the higher-order evidence is defeating for the lower-level beliefs, we know that these norms suggest opposite attitudes: if 1) requires the agent to believe that *p*, then 2) requires, at least, the agent not to believe that *p*. Now I have two norms giving me two different verdicts. The situation seems similar to this: imagine you're in an unfamiliar city, looking for a specific store. You have no access to any sort of map and have no idea where the store is, so you decide to ask for directions to that store. You come across two friends, ask for directions and one of them tells you

to go North, and the other one tells you to head South. They leave without giving you any additional information. What do you do now? There is no non-arbitrary reason to go North instead of South. The puzzle Lasonen-Aarnio is describing is similar to this one: in cases in which rationality gives conflicting verdicts, there seems to be no good reason to obey to one and violate the other. The evidence is misleading because it forces a situation in which you should both believe and not believe a certain propositional content, and it is rationally misleading because the agent is not to be blamed for not being able to follow rational rules that are unfollowable.

1.4.2 Irrationally misleading evidence

Horowitz (2014) argues that views like the ones put forward by Williamson and Lasonen-Aarnio (views which she labels as "level-splitting", because they allow for a mismatch between first and higher-order beliefs) have heavily counter-intuitive consequences. Her point is that if we allow agents to remain epistemically akratic, i.e., if we classify such cases as rational, then it seems that we must also allow agents to engage in other theoretical and practical activities using those akratic beliefs as premises/reasons. But agents that reason or act on epistemically akratic conjuncts are bound to display odd behaviours. She starts her paper with an alleged case of epistemic akrasia:

> **Sleepy Detective:** Sam is a police detective, working to identify a jewel thief. He knows he has good evidence - out of the many suspects, it will strongly support one of them. Late one night, after hours of cracking codes and scrutinizing photographs and letters, he finally comes to the conclusion that the thief was Lucy. Sam is quite confident that his evidence points to Lucy's guilt, and he is quite confident that Lucy committed the crime. In fact, he has accommodated his evidence correctly, and his beliefs are justified. He calls his partner, Alex. "I've gone through all the evidence," Sam says, "and it all points to one person! I've found the thief!" But Alex is unimpressed. She replies: "I can tell you've been up all night working on this. Nine times out of the last ten, your late-night reasoning has been quite sloppy. You're always very confident that you've found the culprit, but you're almost always wrong about what the evidence supports. So your evidence probably doesn't support Lucy in this case." Though Sam hadn't attended to his track record before, he rationally trusts Alex and believes that she is right - that he is usually wrong about what the evidence supports on occasions similar to this one. (Horowitz, 2014: 719)

Horowitz asks us to suppose that Sam has the sort of combination of attitudes that the level-splitters suggest: he believes that Lucy is the thief and believes that he should not believe that Lucy is the thief. Horowitz then suggests that this is the sort of betting behaviour we might expect from Sam:

> Sam: I'd bet that it's Lucy. I'll give you 9:1 odds.
>
> Alex: But you were so sleepy when you were working last night! How can you be so sure that the evidence supports her guilt?
>
> Sam: Oh, I'm not. Since you told me that I'm so bad at evaluating evidence

> when I'm tired, I doubt the evidence supports Lucy's guilt much at all. If I were to bet on what the evidence supported, I might give you 1:9 odds that it's Lucy, but certainly not 9:1.
>
> Alex: So why are you offering 9:1 odds?
>
> Sam: Well, I really shouldn't be offering such strong odds. I shouldn't be so confident that she's the thief: the evidence isn't in my favor. But on the other hand, she is the thief! That's what we're betting on, right? (Horowitz, 2014: 727-728)

Horowitz's point is that because Sam is highly confident of both his beliefs and his beliefs are in conflict, he is incapable of being consistent about the level of confidence he has on those beliefs. When pressed about the first-order belief, he is highly confident of it, and not so confident of the higher-order belief; and when pressed about his higher-order belief, he is highly confident of it, and not so confident of the first-order belief. Another source for suspicion about Sam's behaviour is his suggestion in his final intervention that betting on who the thief is is independent of higher-order considerations. Sam is saying that his betting behaviour should be informed only by his first-order beliefs, and not be affected by any higher-order concerns he might have about his beliefs, even though he has just admitted that he has such concerns!

Furthermore, Sam's odd behaviour won't be limited to bets with his colleague (Horowitz, 2014: 728). If, just like his betting behaviour, his other actions should be informed by his first-order beliefs, then he should do his best to send Lucy to jail. But how will Sam behave in court? He will be called to present the evidence that supports the accusation, but he won't be able to do that successfully because he is convinced that the evidence does not support that accusation at all. Sam's level-splitted beliefs have now led him in a path that could end up with him losing his job. Horowitz, then, concludes that if evidence is misleading, then in those cases in which we let ourselves be misled by it, we are irrationally misled.

1.4.3 Evidence, fragmentation and an approximation

Both Williamson and Lasonen-Aarnio allow evidence to lead to situations in which the agent is not to be blamed if she reaches a state of epistemic akrasia. The main point is that evidence can sometimes detour the epistemic agent, and it is this very characteristic of evidence that softens the stringency of the constraints on epistemic rationality. Horowitz agrees that it is a characteristic of evidence that it can bee misleading, even though she doesn't think that this exempts the epistemic agent. But for all these authors, epistemic akrasia is explained not by some mechanism in the agent's brain (as it was for Greco), but, instead, at the expense of some other factors - in this case, the nature of evidence. The main issue with these views seems to be that they do not provide a clear picture of what is going on in our brains that makes us capable of being epistemically akratic. This is the best feature of Greco's account: the suggestion of a psychological mechanism that explains cases of epistemic akrasia. Just like I noted in the chapter on fallibility centred approaches, this is not a devastating objection to those that propose accounts centred on the nature of evidence, for the views can, presumably, be supplemented with such a mechanism. My point here is rather to note that a satisfying account of epistemic akrasia should include an explanation of what is going on in our brains. However, Greco was subject to

objections whose root may be traced back to his commitment to an explanation based solely on what is happening in our brains. What these chapters seem to suggest is that an adequate explanation of epistemic akrasia should contain not only an explanation of what is going on in our brains, but also an explanation of what triggers that mechanism. This is the direction I shall explore.

Chapter 2

Epistemic akrasia and Constraints

2.1 Epistemic akrasia and practical akrasia

I finished the last chapter by criticizing Greco's view. However, there are some positives to be taken from Greco. One aspect in which his proposal is particularly good, is in its ability to hold that there is some sort of continuity between epistemic akrasia and its practical counterpart. Greco developed his view by analogy with a proposal of characterization of practical akrasia. This enabled him to highlight that continuity and to offer a largely unified explanation of both kinds of akrasia. In this chapter, I'll briefly show how Greco's "construction by analogy" was made, argue for a unified explanation of both kinds of akrasia and show how one can pursue such strategy, while moving away from Greco.

2.1.1 Greco and Gibbard

Greco tells us that his proposal was developed by analogy with Gibbard's (2003) characterization of practical akrasia, and that what he is doing is merely generalizing Gibbard's account to accommodate cases of epistemic akrasia (Greco, 2014: 209). Gibbard characterizes practical akrasia at the expense of a conflict between motivational systems. The main thought is that there are (at least) two different motivational systems, which Gibbard calls the "animal control system" and the "normative control system", and that these may "pull in different directions".

> This is a picture of two motivational systems in conflict. One system is of a kind we think peculiar to human beings; it works through a person's accepting norms. We might call this kind of motivation *normative* motivation, and the putative psychological faculty involved the *normative control system*. The other putative system we might call the *animal control system*, since it, we think, is part of the motivational system that we share with the beasts. Let us treat this picture as a vague psychological hypothesis about what is going on in typical cases of ''weakness of will.'' (Gibbard, 1990: 56)

The distinction between the two motivational systems is made at the expense of a notion of linguistic embedment.

> In many cases a norm is stated in language, or thought in words. A norm, we might say, is a linguistically encoded precept. Perhaps, then, we should think

> of the motivation I have been calling "normative" as motivation of a particular, linguistically infused kind – a kind of motivation that evolved because of the advantages of coordination and planning through language. (Gibbard, 1990: 56-57)

That the normative control system is exclusive to users of languages, ensures that this system is only possessed by fairly sophisticated beings (such as ourselves). At the same time, this allows Gibbard's and Greco's theories to apply to non-human agents: so long as the akratic agent is a linguistic agent, their theories work all the same.

Greco tells us that the notions employed by Gibbard to explain practical akrasia naturally generalize to cover cases of epistemic akrasia (Greco, 2014: 209): the animal control system becomes the non-linguistically infused belief formation system, the normative control system becomes the linguistically infused belief formation system and these belief formation systems "pull in different directions" by arriving at contrasting verdicts, much like what happens (according to Gibbard) in cases of practical akrasia.

> The notion of a linguistically infused belief system that operates alongside a non-linguistic belief system seems just as applicable in the epistemic case as the notion of competing motivational systems is in the moral/practical case. For example, imagine a gambler who is disposed to bet on red if the last 10 spins of the roulette wheel have come up black, unless she says to herself something like the following: "The gambler's fallacy is a fallacy, and the fact that black has come up many times recently doesn't mean that red is any more likely to come up the next time." Suppose that when she does remind herself of these facts, she bets in line with the true probabilities, rather than the ones that would be suggested by the gambler's fallacy.
>
> In such a case, we might distinguish between the dispositions produced by her non-linguistic belief system—call these "beliefs$_n$"—and the dispositions produced by her linguistically infused belief system—call these "beliefs$_l$". We might go on to say that she believes$_n$ that red is more likely to come up after a string of blacks, but at the same time she believes$_l$ that this is not the case, and that the chances of red coming up on any given spin are independent of the results of the previous spins. (Greco, 2014: 209)

Greco's strategy does allow him to stress the continuity between the different kinds of akrasia and to offer a largely unified account of akratic phenomena. This has several advantages. First, it helps justify the application of the same word, 'akrasia', to phenomena in different realms. Second, it helps develop both sides of the theory at once, since someone working on practical akrasia using a theory of type *t* can learn from the mistakes of someone working on epistemic akrasia using a theory of the same type.

2.1.2 A different starting point

What if one wants to follow Greco's strategy but doesn't agree with the conclusions he reaches? One might agree that a unified explanation of different kinds of akrasia is the way to go, but disagree with the idea that epistemic akrasia is to be characterized at the expense of a picture of belief formation that forces one to endorse expressivism about the content of epistemic beliefs and statements. One way to do so is to begin with a different starting point. Greco's starting point is Gibbard's view on practical akrasia. His

commitment to expressivism derives from Gibbard's commitment. In order to get rid of such commitments and to reach different conclusions about epistemic akrasia, while still applying Greco's strategy, is to start by looking at alternative accounts of practical akrasia.

There are many different accounts of practical akrasia. Staying close to another of Greco's intuitions, I'll discuss alternative accounts that also focus on belief formation. I'll start by discussing Davidson's (1969) influential account and advance reasons to reject it, before turning my attention to the account of practical akrasia that I want to focus on.

2.1.2.1 Davidsonian accounts

Davidson starts his discussion of practical akrasia by discussing the meaning of phrases such as 'the agent judges that, all things considered, doing *A* is better than doing *B*'. The puzzling piece of such phrases is the '*all things considered*'. What do we mean by 'all things considered'? What criterion allows us to determine if one particular belief was formed by taking into consideration all things? Davidson's proposal is that we should drop the troublesome 'all things considered' and replace it with a more enlightening phrase. His proposal for a characterization of practical akrasia is:

> An action *x* is incontinent if *x* is done for a reason *r*, and there is no reason *r'* (that includes *r*), on the basis of which the agent judges some action better than *x*. (Davidson, 1969: 40)

That is, an action is akratic if there is no reason that would make the agent change her mind. In other words: an action is akratic when the agent disregards at least one reason that would make her judge some other action to be the best action. The intuition of 'all things considered' is replaced by the more precise 'disregarding no relevant reasons', where a 'relevant reason' is a reason that has some sort of influence over the agent's judgement.[1] This leads to the formulation of Davidson's *Continence Principle*.

> **The Continence Principle:** perform that action that you judge best in light of all relevant reasons. (Davidson, 1969: 41)

Now we have a criterion to determine which actions are akratic, and which aren't. Those actions that are in accordance with the agent's judgement that takes into account every relevant reason, are non-akratic; those that are not, are akratic. Davidson tells us that his Continence Principle is a principle of practical reasoning analogue to a principle of theoretical reasoning introduced by Carnap (1947). Carnap's principle is a commitment of the users of inductive reasoning. It is not a principle regulating the "validity" of inductive arguments, but "a *maxim* that regulates the use of inductive reasoning" (Zilhão, 2010b: 56). This requirement is supposed to prevent abusive applications of inductive reasoning. Given its non-monotonicity, inductive reasoning is susceptible to abusive manipulations: the agent can leave out information from the premisses, and change the probability of the conclusion in accordance with her interests. Carnap's requirement is meant to block such illegitimate applications of inductive reasoning.

[1] Reasons with no influence over the agent's judgement are not relevant for Davidson's purpose. Perhaps an example of an irrelevant reason would help make this point clearer. Imagine that you're trying to decide whether to eat a banana or some strawberries. A proposition such as "The house down the street is being painted" is not a relevant reason because it will not influence your decision whatsoever: your motivation to eat bananas or strawberries is the same whether or not you consider the fact that the house down your street is being painted.

> **The Principle of Total Evidence:** include in the premises of your argument all the relevant available evidence. (Carnap, 1947: 138-139)

In the theoretical realm, a relevant piece of evidence is a piece of evidence that has some influence over the probability of the conclusion. Only those pieces of evidence that do not alter the probability of the conclusion are deemed as irrelevant. The analogy with the practical case is easy to see: pieces of evidence become reasons, and conclusions become judgements about the best course of action; and those reasons that do not influence the judgement are as irrelevant as those pieces of evidence that do not alter the probability of the conclusion.

Davidson claims that both the Continence Principle and the Principle of Total Evidence are principles that the rational agent will observe. And he adds that, even though it might be difficult to make such principles our own principles, that is, to not merely observe them, but to embrace them, this should not be any harder than becoming chaste or brave (Davidson, 1969: 41). I'll now present some reasons to seriously doubt these claims. If I am right, not only is becoming continent incredibly harder than becoming chaste or brave, but, if we stay true to Davidson's account, the whole distinction between being akratic or non-akratic becomes meaningless.

The problem with Davidson's account that I want to stress is that it demands too much from (human) agents. So much so, that it makes it incredibly hard, if not even impossible, for them to be continent. The Continence Principle requires an agent to consider all the relevant reasons when forming her belief about the best course of action. But how does an agent determine which reasons are relevant? Can she determine the relevance of each individual reason without considering every single reason? What kind of procedure would allow the agent to determine the relevance of a reason without considering it? All sorts of principles that I can think of would either be arbitrary ways of stopping the count or would require the agent to consider every reason. An arbitrary way of selecting reasons is determined by the following principle:

> **The 7 Seconds Rule:** take into consideration all those, and only those, reasons that you can think of in the next seven seconds.

This principle clearly allows the agent to consider only a limited set of reasons. But the way in which it does so is, also, clearly arbitrary. What's so special about the next seven seconds? What if, in the proceeding seven seconds, the agent can only come up with reasons to eat pizza tonight, when she is trying to decide whether to accept a certain job offer? The 7 Seconds Rule, even though it grants that the considered set of reasons has a reasonable size, does not guarantee that all relevant reasons are in it. However, every other non-arbitrary principle seems to require the agent to go through every single reason and determine, for each one, its relevance. The general form of this principle is "r is a relevant reason if r instantiates property F", where F is whichever property that makes a particular reason a relevant reason (F might even be context sensitive, that is, it may vary as the deliberation processes/situations vary). A principle of this form forces the agent to go through each and every reason available to him: for how else is she supposed to determine which reasons have F?

You could now ask "How does this talk of 'going through each reason' undermine Davidson's proposal?". The problem with considering every reason is how time-consuming such a deliberation process is. Considering a reason takes time. *Prima facie*, the more reasons you consider, the longer you take to make a decision. Agents like us have a massive

amount of beliefs and desires. So, agents like us have a massive amount of reasons to consider. The conclusion this reasoning leads us to is that agents like us should take an absurd amount of time to reach their conclusions. Consider this example: take an agent with a limited amount of reasons (20) and suppose she takes, on average, 5 seconds to consider each reason. If she wanted to proceed rationally, that is, in accordance with Davidson's proposal, then she would have stopped for 1 minute and 40 seconds before deciding, for example, whether she'll have bananas or strawberries for dessert. If 1 minute and 40 seconds doesn't seem that long to you, notice that this is a very limited agent: no adult human agent has only 20 reasons to consider. And the more reasons an agent has, the longer she will take to rationally reach a conclusion: if she had 354 reasons, she should take 29 minutes and 30 seconds to go through the same decision process. The pessimist moral of the story seems to be this: the more advanced a creature is, the harder it is for that creature to be rational; a simple being, one which only had 1 reason in her repertoire, would have it much easier.

The previous examples, even though they already illustrate the problems posed by Davindson's proposal, really are the simpler, more unproblematic cases. More realistic cases are much more problematic. The deeper problems arise from a) implicit beliefs and b) combination of reasons. Let's start with implicit beliefs.

The problem with implicit beliefs is that agents might have an infinite (or, at least, very large) number of beliefs. And if taking a belief into consideration takes time, then, *prima facie*, considering an infinite number of beliefs would take an infinite amount of time. Explicit beliefs are those beliefs that the agent has already entertained; implicit beliefs are those beliefs that, even though the agent has never entertained, then she can be rightfully described as holding them. One example (I suppose) is your belief that our solar system has less than 37 planets. Probably, and if this is not the case assume it for argumentative purposes, you had never considered this particular belief; but it made sense to say that you believed that our solar system had less than 37 planets (because you believed that our solar system has 8 planets). And just as it made sense to say you had this implicit belief, it also makes sense to say that you believed our solar system had less than 41 planets, and less than 53 planets, etc. What we do in the case of the number of planets of our solar system, we could do regarding many other subjects. You can see how we could get an infinite number of beliefs. And if we can get an infinite number of beliefs, we have an infinite number of reasons which our agent has to go through, which would take her an infinite amount of time. But human agents are finite beings; so, human agents cannot satisfy the requirements imposed by Davidson. So, in light of Davidson's proposal, the distinction between akratic and non-akratic action is an empty distinction. It makes no sense to distinguish between akratic and non-akratic actions, because there is no way in which agents could hope to act rationally. This, I think, is my strongest criticism of Davidson's theory: it shows that it is impossible for agents like us to satisfy its requirements and, consequently, that it is an unrealistic picture of action. However, it is grounded on a debatable distinction between explicit and implicit beliefs. To properly defend this argumentative route, I'd have to enter into the debate about which criteria determine the correct attributions of implicit beliefs. But that is going too far for my current purposes. Instead, let's take a look at the argument from combination of reasons.

The argument from combination of reasons starts from the realization that the previous examples were based on an atomistic conception of reasons, which is not realistic. The simpler examples assumed that to each desire and to each belief corresponded one reason, and that these were all the reasons there were to consider. To make this clear, go back

to the first example (the one with the agent with 20 reasons). It was assumed that these reasons derived directly from her beliefs and desires (suppose, for example, that she had 13 beliefs and 7 desires), and that these did not interact to form new reasons. These examples ignored the possibility of combining the propositional contents of her beliefs and desires to form new reasons. And the reason why this is incorrect, this argument goes, is because considering all reasons is not merely considering each atomic reason. In order to consider *all reasons*, one has to consider molecular reasons, as the strength of reasons varies according to the way in which they are considered. To consider all reasons, an agent has to consider each atomic reason and each possible combination of atomic reasons. And this is problematic because the number of reasons to consider increases dramatically, and, with it, the time an agent should take to make a decision. Take a simple set with 4 atomic reasons. In the simpler picture of combination of atomic reasons[2], you get a total of 15 reasons to consider. Again, 4 is a very, very low number of atomic reasons for an agent like us. So you can imagine just how enormous the total number of reasons for an agent like us would be. And in this combinatorial picture, the number of reasons isn't the only problem: as reasons climb higher in the complexity ladder, so does the time it takes to consider such reasons. *Prima facie*, the more complex a reason is, the longer it takes to consider. So, the combination of reasons poses a double threat to Davidson's proposal. It makes it so that it is (almost) impossible for agents to satisfy its requirements. And even if these are not impossible to satisfy, they certainly are impossible to satisfy *in a reasonable amount of time*.

One way to try to get out of this conclusion while staying true to the spirit of Davidson's proposal is to make slight adjustments to the Continence Principle. These revisions are motivated by Davidson himself (Davidson, 1969: 41 and footnote 24). Davidson suggests that the reader should look at Hempel's *Aspects of Scientific Explanation* for important modifications to Carnap's principle. In this work, as well as later in his "Maximal Specificity and Lawlikeness in Probabilistic Explanation", Hempel introduces a requirement for the users of inductive logic which is weaker than Carnap's:

> **The Principle of Maximal Specificity:** include in the premisses of your arguments all the available evidence that you know influences the probability of the conclusion.[3]

This principle doesn't require the agent to include all the available evidence in the premises of her argument, but only those pieces of evidence that she knows to have some influence over the probability of the conclusion. This principle is more sensitive to the agent's knowledge state than Carnap's. It doesn't blame the agent for not including relevant information if the agent didn't know the information was relevant. It is weaker than Carnap's in the sense that it is less demanding. We can revise the Continence Principle in an analogous way in order to get a weaker principle:

> **The Revised Continence Principle:** perform that action you judge best in light of all the available reasons that you know are relevant.

[2] A more complex picture of combination of reasons would consider different ways reasons can combine. The simpler picture, employed in the example, only considers logical conjunction.

[3] This informal presentation of the principle is based on notes taken during a course on "Philosophy of Science", lectured by António Zilhão at the School of Arts and Humanities of the University of Lisbon, between February and March 2018.

This principle avoids, or, at least, weakens several problems that afflicted the original principle. First, it (probably) avoids the problem with implicit beliefs. Unless we are appealing to an extreme externalist view, an agent cannot know that a certain belief she never entertained is relevant for the matter at hand. Second, it greatly weakens the problems with the number and complexity of beliefs. Since the set of reasons the agent knows are relevant is a proper subset of the set of reasons available to the agent[4], the former is less numerous than the latter and the molecular reasons which can be formed are, in principle, less complex. So the time an agent takes to consider all the reasons that she is asked to consider is expected to be much shorter.

However, the picture of action provided by the Revised Continence Principle is still a very unrealistic one. Here is why.

First, the deliberation processes could still take an absurd amount of time. Imagine that you are very informed about x and that you're trying to make a decision about x. The Revised Continence Principle would still ask you to stop for a very long time before making the decision. Suppose you know of 300 relevant reasons and that you take, on average, 5 seconds to consider each one. According to the Revised Continence Principle, you should spend 25 minutes deliberating before making a decision. Is this a reasonable picture of action, even if we are talking about rational action? It seems an absurd one: if we produced a movie in which the agents behaved like the Revised Continence Principle says they should, we would end up with the most boring movie ever made![5] This picture of action is clearly at odds with our intuitions about the subject and it is very far from capturing the way in which we use the phrase 'rational action'. We classify loads of actions as rational and, nevertheless, no one spends half an hour deciding whether to have bananas or strawberries for desert.

Second, the Revised Continence Principle suffers from a problem that also afflicts the original Continence Principle. These principles imply that the less reasons an agent has, the easier it is for that agent to be rational. In a way, it rewards ignorance: the more ignorant an agent is, the easier it is to be rational - *ignorance is rationality*! Even though, in certain situations, a right decision is more easily made by the fool than by the wise, the claim that there is a necessary connection between an agent's epistemic growth and her difficulty in behaving rationally seems very unlikely. An explanation of action which implies this connection is suspicious.

So, even though the Revised Continence Principle is weaker than the original, it still has some major flaws. We could try to make further adjustments to Davidson's proposal; but I suspect that the time-consuming and rewarding ignorance properties observed in both the original and revised versions are properties that will be present in every version of a Davidsonian continence principle (albeit varying in degree). I'll now present an account of practical akrasia which does not have this negative consequences.

[4] It seems to me that this is an empirical thesis. It is plausible to think, for every agent like us that exists (or has existed, or will exist), if the agent knows something, than the set of things she knows is a proper subset of the total set of that agent's cognitive contents. But it does not seem impossible that these two sets could have exactly the same members (and be non-empty). So, the claim "the set of reasons the agent knows are relevant is a proper subset of the set of reasons available to the agent" is an empirical claim, and not a conceptual one.

[5] On the bright side, Warhol's *Empire* would finally have some genuine competition for the coveted distinction.

2.1.2.2 Fast and frugal heuristics and akrasia

One alternative account of practical akrasia which does respect the requirement that decisions and actions should take place in a reasonable amount of time is advanced by Zilhão (2005; 2010a). Zilhão argues that practical akrasia is explained by a conflict not between motivational systems (as Gibbard maintained), but between belief formation systems. And, unlike Davidson, the conflict between belief formation is not explained by the amount of reasons considered, but by the distinct characteristics of different cognitive systems. He distinguishes between two sorts of judgements: fast judgements and slow judgements.

Slow judgements are the product of explicit processes of deliberative reasoning. Zilhão's slow judgements differ from Davidson's judgements in that Zilhão's slow judgements include any judgement that is the product of an explicit process of deliberative reasoning, imposing no restriction on the underlying inferential strategies or on the proportion of considered relevant reasons.

Fast judgements are the product of fast and frugal heuristics. But what are fast and frugal heuristics? Fast and frugal heuristics are "simple procedures that can be modelled computationally and that are used to search for information, stopping that search and make decisions, and thus solve problems, under limitations of time, knowledge, or computational power or, usually, all of them" (Zilhão, 2005: 204) and are "intended to capture how real minds make decisions under constraints of limited time and knowledge" (Gigerenzer et al, 1999: 5). We can call a creature's set of fast and frugal heuristics its 'adaptive toolbox' (Gigerenzer et al, 1999: 29-31). The analogy with a toolbox allows us to capture two important points: 1) different problems, require different tools; 2) a set of tools with different structures is capable of dealing with a load of different problems, that is, it allows the agent to adapt to the various scenarios the agent is faced with.

> The multitude of simple concepts making up Leibniz's alphabet of human thought were all to be operated on by a single general-purpose tool such as probability theory. But no such universal tool of inference could be found. Just as a mechanic will pull out specific wrenches, pliers, and spark-plug gap gauges for each task in maintaining a car's engine rather than merely hitting everything with a large hammer, different domains of thought require different specialized tools. This is the basic idea of the adaptive toolbox: the collection of specialized cognitive mechanisms that evolution has built into the human mind for specific domains of inference and reasoning, including fast and frugal heuristics. The notion of a toolbox jumbled full of unique one-function devices lacks the beauty of Leibniz's dream of a single all-purpose inferential power tool. Instead, it invokes the more modest but surprising abilities of a "backwoods mechanic and used parts dealer" (as Wimsatt, in press, describes Nature) who can provide serviceable solutions to most any problem with just the things at hand. (Gigerenzer et al, 1999: 30)

Akratic action is explained by the possibility of these belief formations systems (the fast and the slow) reaching different verdicts. In some situations, both systems produce beliefs about what is the best course of action and these beliefs do not coincide. And if in one of those situations, the belief that was produced by the fast belief formation system is the belief that determines action, we have a case of practical akrasia: the agent performs an action that she thinks she ought not to; she thinks she ought not to because

the result of her *explicit* belief formation process is a belief that says that some other action was best. This allows us to capture two of Davidson's important intuitions: that in cases of akrasia, the agent is surprised with herself and struggles to make sense of her own behaviour (Davidson, 1969: 42). This account of practical akrasia supports these intuitions by providing an explanation: the surprise and difficulty in understanding come from the fact that the slow judgement is more present to the conscious life of the agent than the fast judgement, and it is the former that guides the agent's evaluations.

This account does not reward ignorance and does not provide an awfully slow picture of action. First, there is no reason why, on this account, it should be easier to be rational if you are ignorant, because no constraints are placed on inferential strategies. In fact, Zilhão takes akrasia to be a common phenomenon, not a somewhat rare instance of irrationality.

> Fast and frugal heuristics are adaptive mechanisms, which, when triggered in the appropriate contexts, produce effective solutions for the problems the agent is faced with. (...) in the appropriate contexts, fast judgements can play an important corrective and/or preventive role in relation to the explicit deliberative reasoning. That is, frequently, and in the appropriate circumstances, the incontinent action is the objectively rational action, while the continent action would constitute, in those circumstances, the objectively irrational action. (...) So, and contrary to what Davidson maintains, from the surprise and difficulty of the agent to understand herself at the time of the action nothing follows about the rationality or irrationality of her effective behaviour. (Zilhão, 2010a: 325)

Second, not imposing psychologically unrealistic constraints on belief formation allows beliefs to be properly formed and actions to be properly performed within a reasonable amount of time. Characters in a movie who behaved like this account suggests would not stand still for half an hour before acting. The movie wouldn't be freakishly slow paced.

So, we've found a good alternative to Gibbard's characterization of practical akrasia. What does its epistemic analogue look like?

2.2 Constraints on the formation of beliefs

In this section, I'll present my proposal to expand Zilhão's approach to practical akrasia in order to deal with cases of epistemic akrasia. I'll start by motivating the claim that the kinds of constraints at play are not merely temporal or computational, before showing how the expanded theory explains cases of epistemic akrasia. I'll finish by integrating my approach in the program of bounded rationality.

2.3 Constraints on the formation of beliefs

The possible worlds semantics provides us with a good formal tool to represent epistemic operators. Take a subject S, an actual world @, and a set of possible worlds W. S knows that p if, and only if, p is true in all $w \in W$ compatible with S's evidence at @. We can say that what it takes for an agent to know something is for the agent to eliminate all the possible worlds that are not compatible with his evidence. S's evidence is a set of propositions (E), and the compatibility between the evidence and the worlds determines an accessibility relation (R) such that $\forall w\ @Rw \leftrightarrow w$ is compatible with E. But this is highly idealized. In real world situations, an agent doesn't evaluate the entirety of her evidence when forming beliefs. She is under several types of constraints when forming beliefs. Some examples to illustrate this point:

> Paul is talking with his friends. The topic of the conversation is the English Premier League. Fred, one of Paul's friends, asks him which team Paul thinks will take the title in the 2020-21 season. Two seconds after asking the question, Fred insists "Paul?", and Paul promptly responds "Manchester City will certainly take the title back!".

> Christopher is very poor, albeit highly educated. When choosing what to buy, he takes his time to determine which is the best option. However, if you ask him for his thoughts on the Copenhagen interpretation of Quantum Mechanics, he will give you a quick and short answer.

> Francine is trying to determine whether or not argument Υ is valid. But Υ's number of premises is in the order of millions. She is just unable to go through all of them. She does what she can and infers that Υ is valid.

> Emily is a young student of Philosophy. She is also a catholic. Naturally interested in the question of God's existence, Emily has dedicated most of her studies to this subject. She read medievals (Anselm, Bonaventure, Aquinas, Duns Scotus,...), moderns (Descartes, Leibniz, Hume, Kant,...) and contemporaries (Plantinga, Mackie, Swinburne, van Inwagen,...). After critically reflecting on her readings, she concludes that she does not have good reasons to believe in God's existence. Furthermore, she strongly believes she has good reasons to believe in God's non-existence. Nevertheless, she keeps her belief in God's existence.

> Rachel, a middle-aged woman, was sexually abused in her teenage years by a high-school teacher. Today, over twenty years after the traumatic event, she still can't shake the belief that she is not safe, despite having overwhelming evidence that she is in fact safe.

Each of these examples illustrates a different kind of constraint.

Paul's example shows us that our belief formation processes are under time constraints. An obvious way in which we are timely constrained is by our finite existence. But Paul's example shows that tighter time constraints play a role in our lives. Time constraints on belief formation are not only determined by our biological properties, but also by the expectations we are expected to meet in meaningful human interactions.

Christopher's scenario is one in which belief formation processes are under resources constraints. It is not the case that Christopher doesn't have time to consider the question more thoughtfully or that he doesn't have the required intellectual background to seriously consider it. It is not even that he is not interested in the question. The point is that Christopher's financial condition forces him to redirect his resources into getting a better life (financially speaking). One might say that his head is in other things.

Francine's case shows that belief formation is under computational constraints. Supposing that Francine has time to go through all the premises, she is still unable to evaluate the validity of Υ because of its computational complexity. Francine's brain is only capable of dealing with so much complexity (and our brains are very similar to Francine's).

Emily's situation is one in which cultural constraints might be in play. Emily's epistemic state is not a result of a lack of time, computational capacities or other resources. What's affecting her belief formation processes is her cultural background or cultural forces. Cultural elements are forcing her to set aside some of the evidence when forming her beliefs (at least, some of her beliefs).

Rachel's example is one in which psychological constraints play a part. This is the most controversial kind of constraint that I'll discuss. I am not certain that these should be looked at as cases of constraints on the evidence an agent should go through. But it seems as if we can treat these cases as akin to cases of computational constraints: what traumas (in part) do is limit an agent's capacity to go through (certain) evidence. Nevertheless, I am merely interested in the constraints psychological adversities might place on belief formation processes; some authors have defended that Rachel's epistemic state is not only a case in which those processes are constrained but a case in which her belief is justified because it is reliably formed and coherent with her body of beliefs (Freedman, 2006). I don't need to go that far. All I need to ackowledge is that in Rachel's case, her trauma is placing a barrier on her evaluation of evidence, which in turns leads her into a false belief.

What these constraints are doing is limiting the agent's considerable evidence. An agent's considerable evidence is a set $E' \subset E$. This also changes the alternatives relevant to the evaluation of her epistemic state. We now have a revised accessibility relation (R') such that $\forall w \ @R'w \leftrightarrow w$ is compatible with E'.

There is a distinction that is important to make here. I'm taking the constraints to limit an agent's *considerable* evidence, that is, they determine a subset of propositions that the agent is to go through in her belief formation processes. This does not validate a "anything goes" approach. It would if I took E' to be the agent's *considered* evidence. It is crucial that our framework still allows for an evaluation of how good the agent did. That E' is a set of propositions that S should consider allows this: it stills allows the agent to be epistemically irresponsible, by ignoring members of E'.

2.4 Constrained responses and epistemic akrasia

Zilhão's (2010a) explanation of practical akrasia can be expanded to accommodate cases of epistemic akrasia, and provide a unified explanation of akratic phenomena. As the examples of the previous section showed, we are forced to form beliefs under several kinds of constraints. This is precisely the kind of situations in which fast and frugal heuristics are supposed to play a role: the triggering of a fast and frugal heuristic enables the agent to search for information and reach a conclusion even when the conditions to do so are less than ideal. The triggering of these heuristics was what gave rise to a conflict between fast and slow judgements, which in turn was what gave rise to mismatches between an agent's actions and an agent's explicit best judgement. At this level, there is a conflict between beliefs: those that were formed via the intervention of fast and frugal heuristics, and those that came as a result of explicit reasoning. However, this is still not enough to accommodate cases of epistemic akrasia. I need to make the case for the claim that conflicts between different levels of belief can be explained by this mechanism. We can show that it is able to explain such cases by applying it to some examples:

> Sam is a police detective, working to identify a jewel thief. He knows he has good evidence—out of the many suspects, it will strongly support one of them. Late one night, after hours of cracking codes and scrutinizing photographs and letters, he finally comes to the conclusion that the thief was Lucy. Sam is quite confident that his evidence points to Lucy's guilt, and he is quite confident that Lucy committed the crime. In fact, he has accommodated his evidence correctly, and his beliefs are justified. He calls his partner, Alex. "I've gone through all the evidence," Sam says, "and it all points to one person! I've found the thief!" But Alex is unimpressed. She replies: "I can tell you've been up all night working on this. Nine times out of the last ten, your late-night reasoning has been quite sloppy. You're always very confident that you've found the culprit, but you're almost always wrong about what the evidence supports. So your evidence probably doesn't support Lucy in this case." Though Sam hadn't attended to his track record before, he rationally trusts Alex and believes that she is right—that he is usually wrong about what the evidence supports on occasions similar to this one. (Horowitz, 2014: 719)

In this case, Sam's belief about the identity of the thief is the result of a slow, albeit maybe faulted by sleep deprivation, belief formation process. He explicitly goes through the evidence, and takes his time to reach his conclusion. On the other hand, his belief about his own epistemic condition is the result of a fast process: Alex made him aware of his condition, which triggers fast procedures to ensure that a satisfactory response is found in Sam's constrained situation. He does not go through a process of evaluating evidence, but responds in a quick way to his circumstances.

> Matt is extremely afraid of flying. When professional obligations require him to travel (even thousands of miles), he either drives or takes a train. He does not travel overseas. When his friends and loved ones travel by air, he obsessively checks the status of their flights online, and calls them as soon as possible after landing to make sure that they're OK. When asked about all this behavior, he doesn't defend it. Instead, he says things like the following: ''Of course the evidence shows that flying is not particularly dangerous—certainly

> less dangerous than driving comparable distances, but I just can't shake the belief that if I fly, my plane will crash and I will die. What's holding it up there anyway? (Greco, 2014: 202)

Matt's recognition that the evidence shows that flying isn't particularly dangerous comes as a result of an explicit reasoning, whereas his belief that the plain will crash is the result of a fast, unconscious process. Matt is under a psychological constraint as a result of his pathological fear, and procedures in his brain assure that he forms a belief that is in accordance with his pathology. The distinctive nature of the sources of his beliefs explains why they can conflict and why Matt represents an epistemically akratic agent. I'd like to make a remark here. As I said above, I believe psychological constraints are the most controversial ones. However, Greco himself admits that he is not certain whether this is a genuine case of epistemic akrasia. So my own insecurity about the status of psychological constraints is made somewhat less relevant.

> Emily is a young student of Philosophy. She is also a catholic. Naturally interested in the question of God's existence, Emily has dedicated most of her studies to this subject. She read medievals (Anselm, Bonaventure, Aquinas, Duns Scotus,...), moderns (Descartes, Leibniz, Hume, Kant,...) and contemporaries (Plantinga, Mackie, Swinburne, van Inwagen,...). After critically reflecting on her readings, she concludes that she does not have good reasons to believe in God's existence. Furthermore, she strongly believes she has good reasons to believe in God's non-existence. Nevertheless, she keeps her belief in God's existence.

In Emily's case, her belief that she doesn't have good reasons to believe in God's existence is the result of a slow reasoning process, whereas her belief in God's existence is the result of a fast process, her organism's way of responding under the cultural constraints placed (presumably) by her background beliefs and education. Her higher-order belief comes as the result of a process that took her several months, a process of collecting and assessing evidence. On the other hand, her first-order belief is one for which she did no such process. She might describe her own situation as "I don't think I should believe in God, but I still can't shake my belief in God!". Her puzzlement to explain why she can't stop believing in God even though she explicitly believes that she has very good reasons to stop doing so can be explained by the fact that the belief forming process that gives rise to her belief in God is a process of which she might not have consciousness.

2.5 Bounded rationality and epistemic akrasia

We have seen that cases of epistemic akrasia can be explained by appealing to a distinction between fast and slow reasoning, and to the mechanisms that underlie it. But what about the rationality of epistemic akrasia? How should we evaluate akratic states?

The explanation I proposed for epistemic akrasia is one that takes the constraints under which the target beliefs are formed as a crucial part of the explanation. The program of *bounded rationality* is a framework of rationality that takes constraints on the agent as crucial for the evaluation of rationality.

> Broadly stated, the task is to replace the global rationality of economic man with the kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist. (Simon, 1955: 99)

The program of bounded rationality, then, takes into account those constraints that place limits on an agent's access to information and computational capacities. If the examples given in the first section of this chapter are good examples, then all the kinds of constraints identified there classify as constraints admitted by the bounded rationality program, as all place some sort of limit either on the agent's access to information or on her computational capacities. This is a good reason to think that the proposed explanation of epistemic akrasia blends in well with an account of rationality that bounds evaluations to constraints. Furthermore, the appeal to fast and frugal heuristics suggests that this explanation of epistemic akrasia should be included in that specific branch of bounded rationality that explains rationality at the expense of the ecological value of the triggered heuristic. A judgement (and an action triggered by that judgement) formed by a heuristic is rational if and only if the heuristic responsible for that judgement is one that, in most cases, would lead to good results in situations relevantly similar to the one in which the agent is. For our purposes, a belief formed by an heuristic will be classified as rational if and only if the triggered heuristic would, in the majority of the situations relevantly similar to the one the agent is, lead to good results. An epistemically akratic state will be classified as rational if and only if the state is brought about by the triggering of an heuristic that, in the majority of the situations relevantly similar to the one the agent is, would lead to good results.

The notion of 'good results' is vague, but there is a certain purpose to that vagueness. In some cases, the results that matter for the evaluation of a belief will be epistemic results, whereas in other cases the result that matters will be practical. A couple of examples should allow us to see this more clearly.

Suppose you're playing a flash round of a game similar to "Who wants to be a millionaire?". In such rounds, the player must try to get the correct answers in a very short period of time. She does not have time to explicitly go through evidence. Instead, heuristics play a role in assuring that she gets the best answers in the constrained circumstances. To evaluate the value of her beliefs, we don't need to look at what she does with those beliefs, that is, we don't need to also evaluate what actions she performs based on those beliefs. In this situation, all that matters are the epistemic results, that is, whether she gets it right or wrong. All that matters is the truth-value of the beliefs. Her beliefs will be rational if the triggering of the heuristic she used usually leads to good results in situations similar to hers, that is, if it is the way that gets the most correct answers in situations like hers.

In Matt's case (above), the situation is different. To evaluate the value of his belief that the plain will crash, we (should) take into account the practical results of such belief. What we must ask is whether the triggering of that heuristic in situations in which the agent is under some sort of pathological distress usually leads to good results. It is not only a question of the truth of the belief, but of what the agent does with the belief.

The question of whether a particular case of a belief formed by an heuristic is rational is not, therefore, a question that can be resolved merely on *a priori* grounds. In order to determine the rationality of such a belief, we must first determine which heuristic was actually triggered. This step can, arguably, be done merely by looking at the description of the case and at a well established list of heuristics. However, there is a second step that involves the gathering of statistical evidence. In order to determine whether a specific

heuristic usually leads to good results, we must first determine what counts as a relevantly similar situation and then empirically check if the results in those situations were good (and, if we want to be even more rigorous, how good). For this account of epistemic akrasia, attributions of rationality to epistemically akratic states cannot be made independently of empirical studies - they can't be made on a purely conceptual level. That is why I won't give examples of rational or irrational cases of epistemic akrasia. Intuition only goes so far, and, as I lack the database of how good the results of the triggering of specific heuristics are in such and such conditions, trying to guess if a specific, even if hypothetical, case of epistemic akrasia is rational or irrational would go against the spirit of the proposal, as I would be trying to solve on conceptual grounds what I take to be a partly empirical question. What I can do, and have (hopefully successfully) done is to provide the conditions that must be fulfilled for using the classification 'rational' or 'irrational'. Given those conditions, we know what empirical work must be done in order to get at verdicts. Such empirical work, albeit interesting, probably does not belong in a philosophical work.

Conclusion

In the first chapters we put forward some desiderata that an adequate explanation of epistemic akrasia had to fulfil, namely, that it had to include both an explanation of what is happening in our brains that enables us to be in states of epistemic akrasia and an explanation of what triggers the psychological mechanism that enables us to be epistemically akratic. We have shown that an explanation of epistemic akrasia can be given that fulfils such desiderata. The proposed account of epistemic akrasia claims that what enables us to be epistemically akratic is the fact that our brain has different types of belief forming processes, which we labelled as 'fast reasoning (processes)' and 'slow reasoning (processes)'. Furthermore, we showed that what triggers the psychological mechanism that can lead to epistemic akrasia (fast reasoning) is the constraints placed on the agent by the situation in which she finds herself. The fast reasoning processes are, precisely, our organisms' way of dealing with demands under such constraints. Finally, we argued that including this account in the bounded rationality program would enable us to provide conditions for the classification as rational of epistemically akratic states. We concluded that these criteria can be put forward at a conceptual level, but the actual attributions of rationality require empirical work.

Much more would need to be said to support the claim that this is one true account of epistemic akrasia. The list of kinds of constraints suggested is, likely, not exhaustive. In order to get a better version of this account of epistemic akrasia, it would be good to try to get a more exhaustive list of constraints. That would allow us to get a better, faster grip on which cases can be classified as cases of epistemic akrasia.

Also, stronger objections to rival accounts must be given, if we are to claim that this is the only acceptable account of epistemic akrasia. Although I raised objections to rival theories, I did remark in several occasions that the objections were not devastating. The purpose of those first chapters was not to show that all other accounts are necessarily wrong, but rather to determine what characteristics a reasonable explanation of epistemic akrasia should have. The dialectic was: lets determine where others have failed and take those points to be fundamental for the construction of our account. If we successfully tackled those points, then we are in a good position to claim that our proposed account of epistemic akrasia is a good one.

References

Ayer, A. J. (1971) *Language, Truth and Logic*, London: Penguin Books.

Borgoni, C. (2015) ”Epistemic Akrasia and Mental Agency” in *Review of Philosophy and Psychology*, **6**, pp. 827-842.

Borgoni, C. and Luthra, Y. (2017) ”Epistemic Akrasia and the Fallibility of Critical Reasoning” in *Philosophical Studies*, **174**(4), pp. 877-886.

Carnap, R. (1947) ”On the Application of Inductive Logic” in *Philosophy and Phenomenological Research*, **8**, pp. 133-148.

Daoust, M. (2019) ”Epistemic Akrasia and Epistemic Reasons” in *Episteme*, **16**(3), pp. 282-302.

Davidson, D. (1969) ”How is Weakness of the Will Possible?” in D. Davidson, *Essays on Actions and Events*, Oxford: Clarendon Press, pp. 21-42.

Egan, A. (2008) ”Seeing and Believing: Perception, Belief Formation, and the Divided Mind” in *Philosophical Studies*, **140**, pp. 47-63.

Freedman, K. (2006) ”The Epistemological Significance of Psychic Trauma” in *Hypatia*, **21**(2), pp. 104-125.

Gibbard, A. (1990) *Wise Choices, Apt Feelings*, Cambridge: Harvard University Press.

Gibbard, A. (2003) *Thinking How to Live*, Cambridge: Harvard University Press.

Gigerenzer, G, Todd, P. and the ABC Research Group (1999) *Simple Heuristics That Make Us Smart*, Oxford: Oxford University Press.

Greco, D. (2014) ”A Puzzle About Epistemic Akrasia” in *Philosophical Studies*, **167**, pp. 201-219.

Hempel, C. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press.

Hempel, C. (1968) ”Maximal Specificity and Lawlikeness in Probabilistic Explanation” in *Philosophy of Science*, **35**(2), pp. 116-133.

Horowitz, S. (2014) ”Epistemic Akrasia” in *Noûs*, **48**(4), pp. 718-744.

Kearl, T. (2020) ”Epistemic Akrasia and Higher-Order Beliefs” in *Philosophical Studies*, **177**, pp. 2501-2515.

Lasonen-Aarnio, M. (2014) ”Higher-Order Evidence and the Limits of Defeat” in *Philosophy and Phenomenological Research*, **88**(2), pp. 314-345.

Lewis, D. (1982) ”Logic for Equivocators” in *Nôus*, **16**, pp. 431-441.

Miller, A. (2010) ”Non-Cognitivism” in J. Skorupski (ed.)(2010) *The Routledge Companion to Ethics*, London and New York: Routledge, pp. 321-334.

Moore, G. E. (1942) ”A Reply to My Critics” in P. A. Schilpp (1942) *The Philosophy of G. E. Moore*, La Salle: Open Court, pp. 533-676.

Moore, G. E. (1944) ”Russell’s ”Theory of Descriptions”” in P. A. Schilpp (1944) *The Philosophy of Bertrand Russell*, Evanston: Northwestern, pp. 175-226.

Moore, G. E. (1993) ”Moore’s Paradox” in T. Baldwin (ed.)(1993) *G. E. : Moore: Selected Writings*, London and New York: Routledge, pp. 207-212.

Owens, D. (2002) ”Epistemic Akrasia” in *The Monist*, **85**(3), pp. 381-397.

Plato, *Protagoras*, in J. M. Cooper (ed.)(1997) *Plato. Complete Works*, Indianapolis: Hackett Publishing Company, pp. 746-790.

Rorty, A. (1983) ”Akratic Believers” in *American Philosophical Quarterly*, **20**(2), pp. 175-183.

Simon, H. (1955) ”A Behavioral Model of Rational Choice” in *Quarterly Journal of Economics*, **69**(1), pp. 99-118.

Stalnaker, R. (1999) *Context and Content: Essays on Intentionality in Speech and Thought*, Oxford: Oxford University Press.

Williamson, T. (2011) ”Improbable Knowing” in T. Dougherty (ed.) *Evidentialism and its Discontents*, Oxford: Oxford University Press, pp. 147-164.

Williamson, T. (2014) ”Very Improbable Knowing” in *Erkenntnis*, **79**, pp. 971-999.

Zilhão, A. (2005) ”The Pertinence of Incontinence” in *Principia*, **9**, pp. 193-211.

Zilhão, A. (2010a) *Animal Racional ou Bípede Implume?*, Lisbon: Guerra & Paz.

Zilhão, A. (2010b) *Pensar com Risco: 25 Lições de Lógica Indutiva*, Lisbon: Imprensa Nacional-Casa da Moeda.

Zilhão, A. (2017) ”Moore’s Problem” in Beziau, J-Y.; Costa-Leite, A; D’Ottaviano, I. M. L. (orgs.)(2017) *Aftermath of the Logical Paradise*, Campinas: Coleção CLE (vol. 81), pp. 379-399.