Mental Fragmentation

Dominik Berger

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill
2018

Approved by:

Ram Neta

Alex Worsnip

Carla Merino-Rajme

ABSTRACT

Dominik Berger: Mental Fragmentation
(Under the direction of Ram Neta)

My goal in this paper is to offer an account of mental fragmentation. I start out by considering the different phenomena that the notion of mental fragmentation has been used to explain. Then I consider and reject an account of mental fragmentation that is found quite often in the literature, namely that mental fragmentation is what allows an agent to have incoherent attitudes of a certain kind, on the grounds that this account can't explain all the phenomena usually connected to mental fragmentation. In particular, this popular account can't explain why it is that agents who have incoherent beliefs and are mentally fragmented appear to be less irrational for holding the incoherent beliefs than agents who are not fragmented. I will ultimately argue that this can only be explained if we regard mental fragmentation as the result of certain structural features of an agent's cognitive processing.

TABLE OF CONTENTS

Introduction

Mental fragmentation has been appealed to in a number of cases in order to explain several otherwise perplexing phenomena, such as how an otherwise rational agent can have obviously incoherent beliefs.[1] But while many people appeal to mental fragmentation as a solution to several problems, most of these do not (as far as I can tell) offer a genuine account of what mental fragmentation really is. My goal in this paper is to make some headway in this direction. In particular, I will start out by considering two paradigm cases of mental fragmentation and by considering the different phenomena that the notion of mental fragmentation has been used to explain. Then I want to consider and reject an account of mental fragmentation that seems to be found quite often in the literature, namely that mental fragmentation is what allows an agent to have incoherent attitudes of a certain kind. I want to show that on this account, mental fragmentation can't be used to explain everything that needs to be explained in the paradigm cases of mental fragmentation. In particular, this popular account can't explain why it is that agents who have incoherent beliefs and are mentally fragmented appear to be less irrational for holding the incoherent beliefs than agents who are not fragmented. I will ultimately argue that this can only be explained if

[1] For fragmentation and incoherent attitudes, see for example Lewis (1982), Egan (2008), Davidson (1982), (1985), for fragmentation, KK and higher iterations of knowledge, see Greco (2014), for fragmentation and divergence between implicit beliefs and explicit actions see Schwitzgebel (2010) and Quilty-Dunn and Mandelbaum (forthcoming), for mental fragmentation and information access failure, see Egan (2008) and Elga and Rayo (ms a), and for mental fragmentation and logical omniscience, see Elga and Rayo (ms b)

we regard mental fragmentation as the result of certain structural features of an agent's cognitive processing. Given such an account of mental fragmentation, we can make sense of all of theconnected to mental fragmentation. Furthermore, such an account recognizes mental fragmentation as a sub-species of a broader phenomenon: Cases in which our failures of rationality are excused (at least to a certain extent) due to the way our cognitive processing is structured.

The goal of this paper is two-fold: (i) it introduces a phenomenon related to mental fragmentation that has so far been unnoticed in the literature: that mental fragmentation mitigates the irrationality of having incoherent beliefs, and (ii) it tries to spell out in a systematic way what kind of account of mental fragmentation one might give to account for this (and other) phenomena.

Chapter 1 - Mental Fragmentation

Mental fragmentation has been posited in many different areas of epistemology. Egan (2008) for example argues that we are sometimes mentally fragmented. In particular:

> "Rather than having a single system of beliefs that guides all of our behavior all of the time, we have a number of distinct, compartmentalized systems of belief, different ones of which drive different aspects of our behavior in different contexts" (Egan (2008), p.1-2)

Mental fragmentation has been appealed to in order to solve several kinds of problems. Consider for example the following two paradigm examples:

> **Princeton Resident:** A resident of Princeton when thinking about Nassau Street believes that Nassau Street runs North/South and is roughly parallel to the railroads. However, when he is thinking about the railroads, he believes that they run East/West and are roughly parallel to Nassau Street. (See Egan (2008), p.4, taken from Lewis (1982))
>
> **Implicit Racist:** An academic explicitly holds the belief that people of all races are equal and is willing to assert and argue for this belief in academic settings. However, when he interacts with people of color he treats them in an inferior way - he doesn't expect them to have good ideas and often dismisses their comments. (See Schwitzgebel (2010) and Quilty-Dunn and Mandelbaum (forthcoming))

The agents in the two cases display incoherent beliefs: The agent in the Princeton Resident case believes that the railroads run east-west, that Nassau Street runs north-south, and that they are parallel. Likewise, the person in the Implicit Racist case seems to believe both that people of all races are equal and that they are not.

But intuitively it seems that there is more to be said about the beliefs of these agents, besides that they are incoherent. In particular it seems that the agents in the above examples are also mentally fragmented - their beliefs aren't given by one coherent set at all, but rather by different *sets* that are activated in different situations. In the case of the Princeton resident, for example, his beliefs are given by two fragments: One fragment that includes the beliefs that Nassau Street runs North/South and that Nassau Street is parallel to the railroad tracks, and another fragment that includes the beliefs that the railroad tracks run East/West and are parallel to the railroad. The first fragment is active whenever the agent thinks about Nassau Street, and the second one is active whenever the agent thinks about the railroads. In the case of the Implicit Racist, one of his fragments (the one that guides his verbal responses and reasoning) includes the belief that all people are equal, and another one of his fragments (the one that guides his unconscious interactions with people of different races) includes the belief that people of color are inferior. In the rest of this paper I will restrict my discussion to the Princeton Resident case, because this is one of the primary examples of mental fragmentation in the literature. However, I think that everything that I say about this case can also be generalized to other kinds of cases of fragmentation.

I take it that the phenomenon of mental fragmentation that is introduced by the two cases above is something that we are all familiar with - in fact every-day agents are very often fragmented in just such ways. It is an interesting question however why we have the intuition that the agents in the above situations are fragmented rather than merely incoherent. What (explanatory) work is the notion of mental fragmentation doing in these two cases above? It seems to me that there are three things that we want to explain by appealing to mental fragmentation. Throughout this paper I'll refer to these three points as the three desiderata that we should want our account of mental fragmentation to fulfill.

(1) Fragmentation helps to make sense of how agents can hold (obviously) incoherent beliefs in the first place

In the literature many people introduce mental fragmentation in order to explain more easily how agents can hold beliefs and attitudes that are clearly incoherent (see for example Egan (2008) and Davidson (1982), (1985)). For example, it seems hard to make sense of how someone might believe that Nassau Street runs North-South, the railroads run East-West and the two are parallel without noticing the incoherence and thus giving up one of his beliefs. This seems especially difficult if we consider the agent to be otherwise rational. But if we consider that such agents are mentally fragmented, this fact might help us explain how it is that these agents hold such incoherent beliefs. If we assume that the incoherent beliefs are part of different fragments and that only one fragment is activated at one time, then we aren't forced to make the implausible claim that the agent holds all of the problematic beliefs in one fragment (or "at once").

And since all of the incoherent beliefs aren't activated at once, it's easier to see how someone might miss that they have these incoherent beliefs.[2]

(2) Fragmentation allows us to make sense of the very particular (and predictable) *pattern of behavior* that goes along with a certain kind of incoherence

Schwitzgebel (2010) and Egan (2008) point out that if we merely consider the agents above as having incoherent beliefs, we wouldn't be able to capture nicely the predictable pattern in which the agents display the particular beliefs (and the behavior that goes along with the beliefs). For example, we can predict which kinds of beliefs will be displayed and guide the agent's behavior in the Princeton Resident case - if the topic of discussion is Nassau Street, then the relevant beliefs will be that Nassau Street runs North-South and that it's parallel to the railroads. If the topic of discussion is the railroads, the beliefs that she'll display in these situations are that the railroads run East-West, and that they are parallel to Nassau Street. We wouldn't, for example, expect her to start talking about the railroad running East-West in cases when we talk about Nassau Street. So the particular pattern of behavior that is displayed by mentally fragmented agents is that (i) they don't seem to display their incoherent beliefs in the same situations, and (ii) they seem to always display a particular belief in a given kind of situation in a very predictable pattern. And it seems that by appealing to mental fragmentation we can nicely explain why this is so: because the agent's incoherent

[2] The main claim of this desideratum isn't that fragmentation is *required* to explain how agents can have incoherent beliefs, but rather that it can help more easily explain why such agents hold on to incoherent beliefs without revising them.

beliefs are part of different fragments, and the different fragments are activated in different circumstances.

(3) Fragmentation seems to explain why a violation of a coherence requirement in some kinds of cases is *less irrational* than in other kinds of cases of incoherence

There is another important observation in relation to mental fragmentation that hasn't been mentioned in the literature so far. Namely it seems that in a situation in which an agent is fragmented, an agent is *less irrational* for having incoherent beliefs than she would be if she had merely incoherent beliefs without being fragmented.[3] In particular an agent is less irrational in these kinds of cases precisely because she is mentally fragmented, and the fact that she is fragmented *explains* why she is less irrational.[4]

In order to see more clearly what I mean by this, consider again the case of the Princeton Resident: Suppose an agent merely held the three beliefs that Nassau Street runs North/South, that the railroads run East/West and that the two are parallel, then such an agent would be highly irrational. But if we suppose that the agent is fragmented, and that the incoherent beliefs are part of different fragments, then it seems that the agent is much *less* irrational for having the incoherent beliefs (even though the agent still has *the very same incoherent beliefs*). And in particular it is

[3] This point seems to be connected somewhat to the first desideratum - if an agent is less irrational for being incoherent when she is fragmented, this might make it easier to explain why she has these attitudes. This is because in such a case we only have to attribute a little incoherence to the agent.

[4] As I understand fragmentation I think that all cases of fragmentation are cases in which the agent is less irrational for having incoherent beliefs (if these beliefs are in different fragments). Suppose one where to say that this is true only of *some* agents - what is the difference between those fragmented agents who are less irrational and those that are not? I can't think of a further criterion that one might use to discriminate the ones that are less irrational from the ones that are not (other than to appeal to the fact that they are fragmented).

because the latter agent is fragmented that her violation of the coherence requirement seems less bad.

This desideratum relies on the fact that some violations of rational requirements are worse than others. I think this is a very plausible claim, but I don't have enough space to defend it here. Let me just give a small example to motivate it. Suppose someone holds 1000 beliefs which are all (taken together) incoherent in a very subtle way - and in order to be fully rational one would have to give up (at least) one of the 1000 beliefs. This violation of a coherence requirement seems *less severe* than a violation in which a person simply believes P and not-P, and a person who has incoherent beliefs in the first way seems thus less irrational than a person who has incoherent beliefs in the second way.[5]

Since the third desideratum isn't something that's previously mentioned in the literature and because much of my argument in this paper hangs on in, I want to spend a bit of time defending the third desideratum from an objection that one might have against it.

Objection

One direction to put pressure on the third desideratum comes from the following thought. Several writers (Egan (2008), Davidson (1982)) have worried about how we can make sense of agents that have incoherent beliefs - and they seem to think that we can't make sense of incoherent agents at all unless such agents are also fragmented. If this is right then desideratum 3 is in trouble. If all cases of incoherence are also cases

[5] Thanks to Ram Neta for pointing out this example.

of fragmentation (by necessity) it seems implausible to think that being mentally fragmented mitigates (to some degree) one's irrationality - otherwise all cases of incoherence would be (more or less) benign.

I think however that the view that incoherence requires mental fragmentation is implausible for two reasons. Firstly, this view doesn't seem to match very well with our intuitions according to which it seems that every-day agents can be incoherent without also being fragmented. And secondly, I think the idea of a view that doesn't allow any incoherence without positing mental fragmentation might ultimately be unworkable because the very *notion* of fragmentation requires that we can have incoherence within a single fragment. Let's take these in turn.

(1) As an example in which an agent is incoherent without (seeming to be) fragmented, consider the following case that was introduced by Adam Elga:

> **Astrology:** My friend Daria believed in astrology. For example, she thought that because of her astrological sign she was going to be particularly lucky over the next few weeks. That was bad enough. But when I tried to persuade her that astrology is unfounded, I discovered something even worse. I gave Daria evidence against astrology—studies showing that the position of the distant stars at the time of one's birth has no bearing on one's personality or prospects. Daria agreed that the studies were significant evidence against the truth of astrology, and that she had no countervailing evidence of comparable strength. But that was not the end of the matter. "I still believe in astrology just as much as I did before seeing the studies," she said. "Believing in astrology makes me happy." (Elga (2004), p.1/2)

Let us stipulate that believing “P” and “My evidence supports not-P” is incoherent (or at least a *form* of incoherence) and so it is irrational to have these two beliefs together.[6] Then it seems that Daria is a perfect example of a case in which an agent has incoherent beliefs without being fragmented. Since Daria admits in one breath so to speak “My evidence indicates that Astrology is false, but I still believe in it,” it would be very difficult to argue that she *is* fragmented and that her beliefs “P” and “My evidence supports not-P” are part of different fragments. After all - she’s having both beliefs in the same situation.[7]

(2) But even if the case of Daria is not convincing, there is a second problem with holding that any time an agent holds two incoherent beliefs, she necessarily has to be fragmented (and the two beliefs have to be part of different fragments). For it seems that if we accept fragmented agents, we shouldn’t rule out the possibility that agents who are fragmented and have incoherent beliefs become aware of their mistake and rectify the incoherence (a process which we might call *unification* - the process of combining the beliefs given in the two (fragmented) sets back into one set and straightening out the incoherence). But I think the possibility of unifying different fragments that contain incoherent beliefs presupposes the idea that agents can (at

[6] One might perhaps object that in the case of Daria, believing “P” and “My evidence supports not-P” is not incoherent after all. But there are other kinds of incoherence that we might substitute instead - for example one might change the case to one in which Daria is a dialetheist and asserts P and not-P (sometimes at the same time!) I don’t see any reason to deny that in this situation Daria has incoherent beliefs.

[7] The only kind of way we can say that Daria is fragmented in this case is if she’s simultaneously aware of both of her fragments - but this seems to *stretch* the notion of fragments. If she’s aware of both of them - why doesn’t she rectify the incoherence and give up one of her beliefs? I am assuming throughout that a necessary and sufficient condition for two beliefs P1 and P2 to be part of the same fragment is that they are able to inferentially interact with each other.

least temporarily) be incoherent *within* a particular fragment. For how are the beliefs in two fragments F1 and F2 to be unified? The agent will first have to notice that two beliefs P1 and P2 in the two different fragments are incoherent. Suppose that the belief P1 is part of a fragment F1 and that F1 is the fragment that's currently active in the situation the agent is in. Now the agent has to become aware of the other belief P2 that she holds and that's part of a different fragment F2, and she has to become aware that these two beliefs are in fact inconsistent.[8] But one doesn't notice that P1 and P2 are inconsistent unless one can determine the logical relations between the two beliefs (and the two beliefs "inferentially interact with each other" in some minimal sense at least). But in order for P1 and P2 to inferentially interact in this way for the agent to notice the incoherence between them, it seems to me that they have to be part of the same fragment. Thus it seems that on the most natural reading of how a fragmented agent becomes aware of the incoherence and changes his mind, she first has to become aware of her beliefs in different fragments and cease to be fragmented, in order simply to notice that these beliefs are incoherent and to modify the inconsistencies between the different beliefs. But of course this presupposes that one can be (at least temporarily) incoherent within a given fragment. And so it seems that any theory that posits fragments and wants to allow for the possibility of an agent to

[8] I am using the words "inconsistent" and "incoherent" as synonymous throughout.

unify beliefs in different fragments and for that agent to have the possibility to rectify her mistakes, one has to allow that one can be incoherent within a particular fragment.[9]

Of course the above argument doesn't yet show the possibility of agents like Daria who comfortably *sustain* their incoherent beliefs for long periods of time. It might still be true that in any case of unification (and in any case of incoherence more generally) agents feel some pressure to resolve the incoherence as soon as possible. But the considerations above nevertheless seem to me to show two things: (i) If agents can be incoherent for short periods of time, I see no reason why they cannot sustain this state (at least in principle) for longer periods of time - even if there is some (psychological or normative) pressure to resolve the incoherence, and (ii) even if the relationship between incoherence and fragmentation is such as to allow only temporary incoherence, it *still* seems that the third desideratum makes sense. For suppose the agent has incoherent beliefs and is fragmented, but then she ceases to be fragmented and has (temporarily) incoherent beliefs within a single fragment. My intuition is still that the agent moves from a state in which she could relatively comfortably (and only mildly irrationally) hold incoherent attitudes, to a state in which there is suddenly a lot of psychological pressure to revise the attitudes. Furthermore, I have the intuition that the agent in this later state is much more irrational for holding the incoherent attitudes (or in any case she would be much more irrational if she held on to these attitudes in this

[9] Perhaps one might object and say that once one becomes aware of incoherent beliefs one holds in different fragments, one immediately suspends judgement on these beliefs until one has figured out which of the beliefs to keep and which ones to give up. And so one might think that an agent needn't be incoherent within a given fragment after all. But I think this objection misses the point of the argument: My main claim was that *in order to notice the incoherence in the first place,* both incoherent beliefs have to be part of the same fragment. So while it might be true that one's response to noticing the incoherence is to immediately suspend judgment on both beliefs, this still doesn't show that one needn't be incoherent in a single fragment. Thanks to Alex Worsnip and Aliosha Barranco Lopez for helpful discussion on this point.

state and did not revise them). And it seems that this judgement is also something that the notion of fragmentation ought to explain.[10]

So I will (throughout the rest of the paper) take these three desiderata for granted and assume that whatever mental fragmentation turns out to be, it ought to at least explain these three phenomena. So having these three desiderata in mind will help us determine more closely *what* fragmentation really is: In particular, fragmentation will be the kind of thing that best allows us to explain these three desiderata.

In the next section I will look at an account of mental fragmentation that has been supposed in the literature and show how it fails to account for all three of the desiderata (in particular, it will turn out that it fails to account for the third desideratum). Then in the last section I will try to outline a particular kind of approach that I think is required in order for fragmentation to satisfy all of the desiderata.

[10] One might point out that the objection shouldn't be that agents can't have incoherent beliefs without being fragmented, but rather that *rational* agents can't be incoherent without being fragmented. But if this is one's view I see no reason why this should be incompatible with desideratum 3.

Chapter 2 - Fragmentation and Structured Incoherence

Many accounts of fragmentation in the literature associate fragmentation with agents who have incoherent beliefs. But they don't just associate mental fragmentation with agents who have incoherent beliefs - instead they associate fragmentation with agents who display their incoherent beliefs in a particular predictable pattern. In order to see what kind of cases of incoherence are associated with mental fragmentation, consider again how the two agents in the paradigm examples of Section 1 behave: None of them displays their incoherent beliefs at the same time - rather, they display their incoherent beliefs in different circumstances. Furthermore, it seems that there is a very nice and predictable pattern to how a given (fragmented) agent will react in a given situation. Just consider the case of the Princeton Resident again: In cases in which the topic of discussion is Nassau Street, he'll reliably act as if Nassau Street runs North/South and is parallel to the railroads. And in cases in which the topic of discussion is the railroads, he'll reliably act as if the railroads run East/West and are parallel to Nassau Street.

The upshot of these considerations is, I take it, that fragmentation is associated with agents who display their incoherent beliefs in a particular pattern. So this suggests an account of fragmentation, according to which we ought to try to understand fragmentation as the mechanism which underlies (or which explains*)* a particular *kind* of pattern of incoherence. Consider thus the following proposal:

> **Proposal 1:** An agent is fragmented iff she has incoherent beliefs that manifest in a robust, predictable pattern. Fragmentation is that (whatever it is) that underlies or *explains* that robust and predictable pattern.

This is (I take it) a very common proposal in the literature.[11] But how well does this proposal deal with the individual desiderata?

Desideratum 1:

Initially it seems that this proposal has trouble in explaining Desideratum 1 because it associates fragmentation only with a certain kind of incoherence. And so not all cases of incoherence are also cases of fragmentation - just consider the case of Daria from the previous section. But notice that the case of Daria (and similar cases) also doesn't offer as much of a problem in explaining the behavior of the agent as the Princeton Resident case does. What was puzzling in the Princeton Resident case is how someone who is seemingly rational can have such obviously incoherent beliefs - and that puzzling fact needed to be explained. But in the case of Daria it is clear that she is

[11] I take it that such a proposal is the one that Egan (2008) has in mind (see also his quote at the beginning of Section 1). Elga and Rayo (ms a) point out that the beliefs of a fragmented agents can be represented by so-called "access tables" that specify which her beliefs the agent is disposed to act on in given circumstances. If we specify the beliefs of fragmented agents in this way, it seems to me a natural suggestion to get a notion of fragmentation by simply defining it as the kind of mechanism which *causes* these different dispositions. It is naturally to do this also because Elga and Rayo emphasize that having different dispositions in different circumstances is important for considering the agent as fragmented. Elga and Rayo however want to stay silent on the cognitive architecture that underlies these different dispositions and so perhaps might allow that *other mechanisms* than fragmentation can cause an agent to behave in different ways.

irrational. And so we don't need any further explanation of why she holds the incoherent beliefs that she does.

Thus, it seems that the difficult kinds of cases of incoherence that need to be explained are the ones in which the agent appears to be rational and nevertheless hold incoherent beliefs. But it seems that an agent who has two contradictory beliefs P and not-P in the same kind of situation does not seem rational. So the only kinds of cases in which an agent can both seem rational and have incoherent beliefs are the cases in which they display their incoherent beliefs in different circumstances - in other words the cases in which the agent would also be counted as fragmented according to Proposal 1. So it seems that at least in the particular kinds of cases in which we might have trouble making sense of the incoherent beliefs of an agent (and which are therefore the kinds of cases in which Desideratum 1 applies) are the kinds of cases in which Proposal 1 applies and the agent is fragmented. And so we can appeal to the agent's being fragmented to *explain* why they have incoherent beliefs.

Desideratum 2:

Since Proposal 1 gives an account of fragmentation as the mechanism which *explains* the predictable and specific pattern in which an agent's incoherent beliefs are displayed, fragmentation can (according to Proposal 1) do the necessary job of explaining that pattern.

Desideratum 3:

So far then, Proposal 1 looks very promising. However, it seems to me that Proposal 1 struggles to account for the last kind of phenomenon - namely why it is that an incoherent agent who is fragmented is *less irrational* for violating a coherence requirement than an agent who is not fragmented. This is because it doesn't seem that the mere mechanisms that cause agents to display her incoherent beliefs in different circumstances are the right *kind of thing* that could explain why her failure to obey the coherence requirement is less irrational. And so it is possible that there might be cases in which someone counts as fragmented (according to Proposal 1), but in which the agent's incoherence is *not* mitigated. For an example of such a case, consider a variant of the Astrology case:

> **Astrology with Friends:** Daria still believes that Astrology is true and that her evidence doesn't support astrology (because believing in Astrology makes her happy). But now suppose that she doesn't assert both things at the same time - instead she asserts each claim in different situations. She asserts that Astrology is true whenever she spends time with her Astrology friends - strangely the question about whether the evidence supports Astrology doesn't really come up in this context (and so Daria never thinks about the evidence there). Additionally, Daria has a lot of Philosopher friends who talk a lot about what the evidence supports and who also never question whether Astrology is actually true. So unsurprisingly, Daria never thinks about her belief that Astrology is true in this context.

Let's suppose that in this kind of case, Daria does in fact end up asserting that Astrology is true mostly in particular kinds of situations (when she's alone and when she spends time with her Astrology friends), and asserting that the evidence doesn't support Astrology in some other kinds of situations (when she's with her Philosophy friends). But it's not for any *deep* reason - it's just that whenever she's by herself and with her Astrology friends, the question of whether the evidence supports Astrology doesn't come up and so she never thinks about it. And similarly the question of whether anyone believes in Astrology doesn't come up when she's talking with her Philosopher friends, and so in these situations too, she never thinks about Astrology. In these kinds of circumstances, then, she will robustly assert that P in one kind of situation, and that her evidence doesn't support P in some other kind of situation. In other words, Daria has "incoherent beliefs that manifest in a robust, predictable pattern" - and so it seems that we should count her as being *fragmented* according to Proposal 1.[12]

I think it is the wrong result to regard Daria as fragmented in the case above - my intuition is that Daria is simply not fragmented, and that Proposal 1 is thus false.[13] But suppose we were driven by theoretical reasons to accept Proposal 1 and are therefore tempted to count Daria in this case as fragmented (in line with Proposal 1 but contra intuition). Then one also gives up the aspiration that fragmentation can explain why it is that an agent who has incoherent beliefs (but is not fragmented) is more

[12] You might think that this example is also a problem for Proposal 1 on another count. For remember that Proposal 1 defines fragmentation as the *mechanism* that underlies (or explains) that the agent displays her incoherent beliefs differently in different circumstances. But you might think that in the case of Astrology with friends, what explains that she displays these beliefs in different circumstances isn't a fact about her mind and about how her beliefs are organized, but rather about her environment. I take it that this is also a nice point that should worry someone who accepts Proposal 1. But I'm not convinced that this poses an inescapable problem - perhaps one might say that the way her environment is structured imposes a certain kind of organization/structure on the agent's beliefs and her dispositions to display these beliefs, and this *organization* is what explains her behavior (and it is this structure that is thus picked out by Proposal 1 as counting as mental fragmentation).

[13] One might worry here that this case is very similar to the case of the Princeton Resident and so think that if the Princeton Resident is fragmented, then so is Daria. But I think we should resist this temptation. I will come back to this worry in the next Section after proposing my own account of mental fragmentation. This account will help us see more clearly, I think, what the difference is between Daria and the Princeton Resident.

irrational than a fragmented agent with incoherent beliefs. This is because Daria in the case above doesn't seem to be any less irrational for having the two incoherent beliefs "P" and "My evidence supports not-P" than she was before. For what made it the case that she was irrational in the previous case wasn't just that she asserted her incoherent beliefs in one breath - it was that she had those incoherent beliefs and was aware of them (in some sense). So it's the having of those incoherent beliefs that makes her irrational - and merely displaying different dispositions in different circumstances by itself isn't the kind of thing that could make it the case that she has become less irrational.[14] So if we are following Proposal 1 in calling the agent fragmented in this case, then fragmentation loses the ability to explain (by itself) why incoherent, fragmented agents are less irrational than incoherent, non-fragmented, agents. For recall the intuition behind desideratum 3 is that the *mere fact of being fragmented* (and for one's incoherent beliefs to be part of different fragments) suffices to mitigate one's irrationality for having incoherent beliefs.

Objection:

One might say that I've made a mistake in spelling out the case of Daria in this way, because if someone were to ask Daria "Do you believe in Astrology" she would say yes - in both kinds of situations. And similarly, if anyone were to ask her "Do you believe the evidence supports Astrology" she would say no - in both situations. And so it doesn't seem that all Daria's dispositions in one situation can ever be completely pro-Astrology (or anti-Astrology).

[14] Thanks to Aliosha Barranco Lopez for suggesting to me this way of making the point clearer.

But notice that it would be a mistake to count her dispositions or behavior when *explicitly prompted/asked about one of the attitudes* as making the relevant difference here to whether we should consider her as fragmented or not. For after all, the Princeton Resident when asked about the direction of Nassau Street will also always say “It runs North-South” and when asked about the direction of the railroads will say “They run East-West” - and so if their willingness to answer questions when explicitly prompted counted among the relevant dispositions, then *no one* in the original examples would really count as fragmented. So this seems that the dispositions that we need to take into account with regards to Proposal 1 are the agent’s dispositions to display certain behaviors or to have certain beliefs in the original cases *without* interfering with the agent. Once we explicitly ask them about a given piece of information it seems that we will change how (or whether) the agent is fragmented.

So I think that the case of Astrology with Friends shows that there can be cases in which an agent is fragmented according to Proposal 1, but in which the irrationality for violating the coherence requirement is not mitigated. This shows, I think, that on Proposal 1 fragmentation can’t explain (at least not by itself) why it is that a fragmented agent is less irrational for violating a coherence requirement than she would be if she were not fragmented. In other words: it doesn’t seem that fragmentation is the thing that *makes the difference* in determining how rational she is.

But even if you are not convinced by the Astrology with Friends example (either because you think it’s not a genuine case of fragmentation according to Proposal 1, or because you think that in this case Daria is less irrational for having incoherent beliefs) and it turns out that Proposal 1 considers agents fragmented iff they really are fragmented, it still seems to me that we are missing a crucial ingredient in our account of fragmentation. This is because appealing merely to the mechanisms that cause an agent to display her incoherent beliefs in different kinds of circumstances isn't *the right kind of thing* that could ever *explain* why fragmentation makes one less irrational in all these kinds of cases. Suppose (for example) we should actually consider Daria in this example as a case of genuine mental fragmentation - why should we consider her as less irrational? It doesn't seem that this kind of proposal by itself could offer a significant answer. So even if Proposal 1 doesn't make false predictions about when we should regard agents as fragmented, we still need to supplement it with *something else.* Thus, it seems that having an agent that is (i) incoherent, and (ii) the incoherence is displayed in different situations is not enough to account for these agents as being fragmented. It seems that we need to refine our account of fragmentation and add something else. In the next section, I'll have a look at what this something else might be so that we could *explain* why an incoherent agent is less irrational if she is fragmented.

Chapter 3 - Mental Fragmentation and Cognitive Processing

In summary, then, I have shown that an initially attractive account of mental fragmentation fails. Furthermore, it fails because it can't be used in an explanation for why we think that incoherent, fragmented agents are *less irrational* than agents that are merely incoherent.

So in this section I want to investigate in more detail what kind of explanation we might give in order to explain this judgement. In order to do so, let us first look at other cases that do not concern fragmentation in which it seems that an agent's violation of a rational requirement is somehow mitigated*.* One such example is the case of logical truths and logical consequence.

Logical Consequence and Irrationality

Consider, for example, the following kinds of mistakes one can make about logical truths and logical consequences:

(1) **Simple Closure Failure:** Suppose a person believes P and Q which entail R - but the person doesn't believe R

(2) **Complicated Closure Failure:** Suppose a person believes P1, …….,Pn (for some large n) all of which together entail R, but the person doesn't believe R

Suppose that in neither case the agent is fragmented. Then it seems at least intuitive that an agent who is making a mistake of the second sort is less irrational for making that mistake than for making mistakes of the first sort. In general it seems that failing to believe logical truths or logical consequences of one's beliefs is *less* irrational the *more complicated* the logical truths or logical entailments are. Let us reflect a moment on why that is.

It seems to me that the problem isn't necessarily that the specific truths (or the particular logical entailment relations) are very complicated. Consider, for example, one particularly complicated logical truth T. Failing to believe T isn't as irrational as failing to believe a simpler logical truth (let's call it S) - but this isn't a property of T, such that making a mistake with respect to T is less irrational than making a mistake with regards to S for *every agent*. For just consider some alien creatures that are (nearly) perfect at logic and who compute and comprehend all logical truths and logical entailments as easily as we understand 2+2=4 (or at least logical truths that are far more complicated than S or T). For such creatures, making a mistake with respect to T is *just as irrational* as making a mistake with regards to S.

This observation suggests that what makes it less irrational to make a mistake with regards to T rather than S isn't a fact about S and T, but rather a fact about us. More specifically - a fact about how we compute or think about logical truths. In other words: a fact about our *cognitive processing.* It's much harder for us to compute difficult logical entailments because we can't (for example) keep several steps of a long logical derivation in our minds at the same time. So computing a difficult logical consequence might require a lot of time and mental energy (or the help of pencil and paper, etc.) If, on the other hand, our cognitive processing mechanisms were built differently in such a way such that we were better able to compute logical truths (like in the case of the aliens), then it would not be less irrational if we made a mistake about more difficult logical truths. This is why for the aliens it is just as irrational to make a mistake of the first sort as it is to make a mistake of the second sort.

Fragmentation and Belief access

I want to suggest then, that something similar could explain why it is that incoherent agents who are mentally fragmented don't seem as irrational as other agents who are not. The proposal, I think, will be something along similar lines as the one proposed above for the case of logical truths and entailments: Fragmentation acts as a mitigating factor for agents with incoherent beliefs just because it is somehow *difficult* to access information stored in different fragments together (and so notice the incoherence). But it's not just that it's difficult - it's difficult *because of the way our cognitive processing works.* That is, I want to consider the following proposal:

> **Proposal 2:** An agent is fragmented in situation S just in case some of the agent's beliefs in situation S are difficult to access given the way her cognitive processing works from situation S. Fragmentation is the phenomenon of not having easy access to some of one's beliefs in a given situation due to the way one's cognitive processing is structured.

It seems to me that Proposal 2 is on the right track. But it isn't really complete without a nice account of what it means for a given belief to be difficult (or easy) to access in a given situation. Unfortunately I don't quite have a general account of what it means for some beliefs to be easily or difficult to access. It seems to me that a fully satisfying account will point out the particular cognitive process that's employed in the Princeton Resident case (and other cases of fragmentation) in order to access one's beliefs and point out how for this process some information (the one in the Nassau Street fragment) is easily accessible, and some *other* kind of information (the one in the Railroad fragment) is difficult to access.

Instead of giving such a general account (or specifically pointing out the relevant cognitive processes), I just want to point to three examples of cognitive processes from Cognitive Psychology and Cognitive Science to suggest that there really is a distinction between beliefs that are easily and difficult to access, and that this distinction depends on the cognitive processes we use to form beliefs, and that which beliefs are easy and difficult to access will be different in each situation. I want to leave it open how precisely the details are going to be spelled out in each particular case of fragmentation.

Example 1: Semantic Priming

The idea behind the phenomenon of Semantic Priming is this: Subjects are asked to classify strings of letters (such as "street" or "salfem") into "words" and "non-words". In some situations, the subject is *primed* with a particular kind of word - in other words: they are shown a particular kind of word for a little bit before being asked to classify a *different* word as word/non-word. In cases in which subjects are primed with a word of a particular category (say "doctor"), participants are much faster at identifying words in a related category (such as "nurse") as words than they are at identifying words in a different category (say “sandwich”) as words. The case of semantic priming gives us an example of a case in which some information is *easier to access* than other information due to facts about the way our cognitive processing seems to be organized.[15]

Example 2: Heuristics

It seems that a lot of our cognitive processing depends on heuristics - which are rough and quick rules of thumb according to which we reason. Part of what makes these rules of thumb so effective is that they only consider a very small and select subset of information and ignore all of our other beliefs. So if we are in situations in which we’re employing such a rule, then it seems that once the given cognitive process is initiated, all the beliefs that are in the domain of the rule (i.e. that are the kinds of beliefs that the rule typically has access to) are easily accessible, and all other kinds of beliefs (the ones that are *not* in the domain of the heuristic/rule) are difficult to access.[16]

Example 3: Modular Processing

It seems that many of our mental processes might be modular - in other words, it seems that many of our processes are fast and automatic and take into account only a small subset of all of our beliefs.[17] If a particular action (or belief) is the output of a modular process, it seems that in the situation in which this modular process is running, the beliefs that are being accessed by this process (or the beliefs that are in

[15] I take the description of semantic priming from Solso, Maclin and Maclin (2008)

[16] For more information about our use of heuristics in order to process information, see Kahnemann (2013)

[17] I rely on an account of “modular” that is spelled out by Carruthers (2006a), (2006b). Carruthers defends the thesis that all of our cognitive processing is modular in this way.

the *domain* of that process in situation S) are the ones that are easily accessible*,* and the other beliefs that are not in the domain of the modular process are difficult to access.[18]

It is not quite clear whether the three processes I have outlined describe different *kinds* of cognitive processes, or rather whether the examples I've described all (more or less) pick out the same kind of phenomenon - namely that in most situations, in order to react within an adequate time-frame to a given situation we only consult a very small subset of our beliefs.[19] Additionally, it seems that each situation comes (roughly) with a guide to *which* beliefs are relevant and which are not - we might call the former *active* beliefs and the latter *inactive* beliefs. I take it that roughly this distinction between active and inactive beliefs is the explanation for why some beliefs are *easily* accessible and others are not - namely the active beliefs are easily accessible. One important thing to notice, however, is that which beliefs are active and which are inactive seems to be not only relative to a particular situation, but primarily relative to the *particular cognitive process* that is activated in the situation.

So I roughly take to be the explanation of fragmentation that was outlined in Proposal 2 to be this: Suppose an agent employs a particular process (such as a heuristic, a modular process, or is primed a certain way, etc.) in a given situation, and suppose further that the given cognitive process is structured in such a way so that it can access only a *small subset* of the agent's beliefs - in such a situation the agent will

[18] For more information about modular processing, see Carruthers (2006a) and (2006b).

[19] This is also called the "frame" problem. See for example Dennett (1984). On some connections between a similar kind of problem and mental fragmentation, see Cherniak (1983).

turn out to be fragmented. And the fact that we can't easily access some of our beliefs is a factor of how our cognitive processing is structured - in particular, it's a consequence of the fact that (i) we often employ heuristics, modular processing, etc. and (ii) these processes can only access a small subset of our beliefs at a time.

Some examples: Princeton Resident and Astrology with Friends

Now, in order to better understand how my proposal works, let us see how it applies in the Princeton Resident's case. We said before that his beliefs are roughly given by two sets. So let's also say that the fragmented beliefs are indexed to two kinds of situations: There is one kind of situation in which the agent is thinking (or talking) about Nassau Street - in this situation the belief that Nassau Street runs North-South and that the Railroads are parallel to Nassau Street are easy to access. But it's *hard* to access the belief that the railroads go East-West because it is inactive. In particular it is difficult to access this belief from the particular situation that the agent is in because this belief doesn't seem to be in the *domain* of the cognitive processes that the agent employs in the given situation. (Roughly I take it that this is so because beliefs about the railroad are deemed irrelevant for talking about Nassau Street.)[20] This is because in cases in which he's talking about Nassau Street, the beliefs about Nassau Street (that it runs North-South and that it's parallel to the railroads - these are after all just facts *about* Nassau Street) are active, whereas the fact about *other things* (that are not Nassau

[20] I should be a bit careful here: Which beliefs are in the *domain* of a particular cognitive process might vary with the situation. Thus we might think that which beliefs are in the domain (and hence easily accessible) is a function of both (i) the cognitive process employed, and (ii) the situation one is in. (The second component will likely give important clues for which information is relevant.)

Street - like the direction of the Railroad) are not active. (And of course we can tell the parallel story involving beliefs about the railroad.)

We can now also see why Daria in the case of Astrology with Friends is *different* from the case of the Princeton Resident and why she - even in cases in which she displays nicely compartmentalized behavior - should *not* count as fragmented. This is because even in cases in which she is spending time with her astrology buddies and displays no dispositions related to her belief that the evidence doesn't support astrology, this belief of hers is nevertheless *easily accessible* from the epistemic standpoint that she occupies at the same time.[21] So Proposal 2 is an improvement on Proposal 1 and is able to get the right result in the case of Daria, because it doesn't count *all* of the agents dispositions to display her beliefs in different situations as relevant to whether the agent is fragmented. Instead, it only counts as relevant the dispositions that are due to some facts about the *structure* of her cognitive processing.

In order to get clearer on the distinction between Daria and the Princeton Resident I want to introduce a very loose and metaphorical way of speaking about the difference between active and inactive beliefs: Let us say that beliefs that are inactive are *far from the agent's mind.* One way to notice when a belief is inactive (as opposed to active) might be in the difference of how the agent reacts when one suddenly brings up this belief: If a belief is truly far from the agent's mind, then she will be surprised and a little bit startled after the belief was brought up (or in any case the agent's reaction will be

[21] Note here the important distinction between a belief that is (consciously) entertained and a belief that is active - a belief might be active even though it's not consciously entertained.

“Huh. Why do you do bring up this piece of information? I hadn’t really thought about it at the moment.”)

So then the difference between Daria and the Princeton Resident is that in the case of Daria, even though she doesn’t think about the state of the evidence when she talks with her astrology friends, this fact is not *far from her mind.* But in the Princeton Resident case, the orientation of the railroads is far from the agent’s mind. For example, if you were to point out to the Princeton Resident (forcefully) that he *did* have this belief about the direction of the railroad (that it goes East-West) while he’s considering the direction of Nassau Street, he would be astonished and a bit surprised. (And perhaps think “Oh wow. Yes, you are absolutely right. I hadn’t considered that at all - I do have this belief!”) But if you were to point out to Daria that she believes that the evidence doesn’t support her belief, she would respond “Yes I know” and *not* be surprised or startled in the slightest - even in cases in which she talks to her Astrology friends. I think this is then roughly the difference between the case of Astrology with friends (where the agent isn’t actually fragmented), and the case of true fragmentation (like the case of the Princeton Resident). In the former case (but not in the latter) even though the incoherent belief isn’t *occurrent,* it is *not far from her mind* (and hence not difficult to access).

Here is another point that suggests that this is *really* what the difference is between the two kinds of cases. Suppose the belief about the status of her evidence is actually difficult to access (and so far from her mind) in situations in which Daria spends time with her Astrology friends, and she would genuinely be surprised if you pointed it out to her. In this case it does seem that my definition would suggest that she is fragmented. But notice that in this case I also have the intuition that *she would be less irrational* for having both of the incompatible beliefs and so that she would in fact be fragmented! And so if the agent satisfies Proposal 2 then it doesn't all of a sudden seem like the wrong verdict at all if in this case we consider the agent as fragmented. This seems like good news for the proposal.

Finally, let us ask how this proposal deals with the three desiderata that I've outlined at the beginning of the paper.

Desideratum 1

Can my proposal explain desideratum 1 - explaining how (seemingly rational) people can have incoherent beliefs? If seems to me that it can. If these incoherent beliefs are part of different fragments in a given situation, then the incoherent beliefs are hard to access at the same time and so it's understandable (and excusable) that the person might not notice the incoherence - even if she is perfectly reasonable. This is because it's difficult (given certain facts about her cognitive processing) for her to do so.

Desideratum 2

How does my proposal deal with the second desideratum - that agents with incoherent beliefs that are fragmented will behave in a nice and systematic way in different situations? This also seems to be easily explained. Suppose you say that what causes a belief to be active in a given situation are facts about the situation (e.g. what kind of action is required, what kind of information is salient, etc). Then it seems that in situations of the *same* kind the *same* beliefs will be active and inactive - and so agents with incoherent beliefs will display these beliefs in a nicely and predictable kind of way. Furthermore it is easy to explain why the incoherent beliefs are always activated in *different* situations and never at the same time - if they were activated at the same time, the agent would notice the incoherence and adjust for it. Thus it seems that what explains the persistence of incoherent beliefs is that they are *not* activated during the same process/at the same time.

Desideratum 3

Similarly it seems that my proposal can accommodate desideratum 3 - why incoherent agents that are fragmented appear to be less irrational than other agents. This is because in the former (but not in the latter) case, it is difficult (given one's cognitive processing) to access all the relevant incoherent beliefs at the same time. And it is the *increased cognitive difficulty* of recognizing one's incoherent beliefs in the case of fragmentation that explains why agents are *less irrational* for not noticing it. (Just like the increased difficulty of computing complicated logical entailments that mitigates one's irrationality if one fails to have such beliefs).[22]

[22] There is an interesting question whether the relevant factors in determining how irrational someone is are facts about the structure of the cognitive processing of the *individual agent* or facts about the cognitive processing of *human agents in general.* How we spell out this proposal might influence whether we think (for example) a child is less irrational for failing to notice certain inconsistencies in her beliefs, or for whether a person with generally poor cognitive processing is less irrational for failing to satisfy some rational requirements. If we embrace the *first* view then they do count as less irrational, but on the second view they don't. I want to leave it open however which way we want to spell this out (though I am leaning towards the second proposal). Thanks for Minji Jang for bringing up this difficulty.

Thus it seems that my proposal can nicely account for all of the three desiderata that we want mental fragmentation to explain - and so it offers a superior account of mental fragmentation than the competing proposal discussed in Section 2.

For all these virtues this new proposal has, it does, however, conflict with the pre-theoretic notion of fragmentation one might have. Suppose one thinks of fragmented agents as agents whose beliefs are broken up into several sets - and while these sets might be activated in different situations and updated as one learns new information, they do stay roughly the same throughout the agent's lives. Then one might not like my account, according to which fragmentation occurs (roughly) whenever one has difficulty accessing some of one's beliefs (which occurs most likely quite often). So in order to show that this skepticism isn't warranted, I want to briefly address three points of criticism that one might make against my proposal.

(i) Suppose fragmentation isn't a question of which information it is difficult for an agent to access in a given situation, but rather a question of having one's information stored in different compartments or blocks (and one doesn't notice information in compartments that aren't activated).[23] While this picture might initially match more closely with one's pre-theoretic notion of fragmentation as being a robust phenomenon of fragments enduring over time, it can't really be quite right. For suppose we designate two stable "compartments" as fragments F1 and F2 - it seems that there could be situations in which both information from F1 and F2 is relevant and in which both information from F1 and F2 is easily accessible due to the cognitive process one

[23] This is roughly the proposal outlined in Cherniak (1983)

employs in a given situation and beliefs from both F1 and F2 inferentially interact with each other. But then it seems difficult to say that the agent is still fragmented (or more accurately: that F1 and F2 nevertheless constitute different fragments).[24]

(ii) One might also be worried that my account of fragmentation simply identifies the phenomenon of fragmentation with failures of information access more generally.[25] But this need not be the case - my account only counts *some* cases of information access failure as fragmentation.

In order to see this, consider for example an agent who *forgets* a piece of relevant information. I want to distinguish two kinds of ways in which one might forget a piece of information. Suppose the agent "forgets" a piece of information P in the sense that one merely doesn't take P into account in a situation in which it is relevant. (Presumably because P is not in the domain of the cognitive process employed in the particular situation that the agent is in.) This situation seems to me to be a paradigm case of an agent who is mentally fragmented. But notice what's important in this situation is that even though P is relevant in the situation and the agent doesn't employ

[24] A similar worry is also brought up in Norby (2014), who argues against fragmentation on the grounds that it does not correspond to anything psychologically real. He considers an account of mental fragmentation according to which the different mental fragments might be the different groupings in which our beliefs are stored together - and rejects it on roughly similar grounds as the ones I brought up above and concludes that fragmentation doesn't correspond to anything psychologically real. But notice that my proposal does identify fragmentation as a psychologically real phenomenon - namely the fact that some beliefs are easily accessed by some cognitive processes and others are not.

Norby mentions a second concern, namely that the mental fragments are explanatorily useless unless we give an account of when a particular fragment is active. He points out that "there simply is no notion of what the *activation* of a belief or belief-fragment is other than that it consists in that belief (or the beliefs in the relevant fragment) guiding one's behavior." (Norby (2014), p. 7) And he argues that without a different definition of activation one is forced to conclude that on the fragmentation picture, one beliefs P (in some fragment) iff one's behavior is guided by P in some situation - which is clearly false. Notice that my proposal can avoid this second problem as well and give a nice account of when a belief is active - namely when it is in the domain of the particular cognitive process that the agent employs in a particular situation.

[25] Thanks to Alex Worsnip and Carla Merino-Rajme for pointing out this objection.

this information, it’s not really a case of *forgetting.* It’s merely a case in which the agent fails to bring to bear a piece of information that is relevant.

Consider a second kind of case that is much more like traditional cases of forgetting however. Suppose, for example, that the agent is trying to access a given piece of information that she *knows* she possesses, but is merely struggling to access it. Suppose for example you ask me what the capital of the Netherlands is and I *know* that I have this piece of information - but I just can’t remember it in this particular moment.[26] In this case it is true that the agent doesn’t seem as if she is fragmented. But I also think that my account need *not* imply that the agent is fragmented. This is because even though it might be difficult for the agent to access the particular piece of information, this difficulty needn’t be due to the *structure* of the cognitive processing that the agent employs. It might very well be that this piece of information is in the domain of the cognitive process that the agent is employing and the cognitive process is trying to access that particular piece of information - but it simply can’t because of some *other* kind of malfunction. If this is the case, then my proposal doesn’t imply that the agent is fragmented.

(iii) There is a last kind of objection one might raise. In this paper I haven’t talked very much about the Implicit Racist case. In Section 1 I brought up the Implicit Racist case as an example of fragmentation - but one might worry that my Proposal might make it in fact quite difficult to see *how* the Implicit Racist case is a case of fragmentation. After all we might imagine the following situation: Suppose that the implicit racist is having a discussion about racial equality with a group of people, and

[26] I owe this kind of example to Chris Blake-Turner.

while she’s vigorously defending the view that people of all races are equal, she also constantly cuts off people of color. It seems that in this case we should still say that the Implicit Racist is fragmented - but it is not necessarily clear that any of her beliefs are *difficult to access* in the situation. After all: She’s actively accessing her egalitarian beliefs in order to defend the view that people of all races are equal, and she’s accessing her racist beliefs when she’s cutting of people of color. So it might be that at least in a situation in which she’s cutting of people of color, her egalitarian beliefs are still on her mind (because she’s in the process of defending them) and so can easily be accessed.

But I think we can avoid this objection if we notice that in that particular case the agent is undergoing two processes at once: (1) She’s voicing her support for egalitarian beliefs, and (2) she’s treating a person of color terribly. On the conscious level she’s undergoing a process that is involved in her supporting her egalitarian beliefs. And it is this process which has (easy) access to her egalitarian beliefs but only very difficult (or no) access to her subconscious racist beliefs. Likewise, another (unconscious) process guides her behavior - and this process has easy access to her racist beliefs but doesn’t have any (or in any case no easy) access to her egalitarian beliefs. So rather than denying that the Implicit Racist is fragmented, I think we ought to say that she is actually fragmented twice over: Once on the conscious level, and once on the unconscious level!

Conclusion

In this paper I have looked more closely at the phenomenon of mental fragmentation, outlined three desiderata that we take this phenomenon to explain, and argued that the most popular account so far hinted at in the literature can't explain all three. Then I have outlined my own account which can explain them in a satisfactory way - namely if we tie mental fragmentation to the phenomenon that often many of our beliefs are difficult to access (given the way our cognitive processing is structured) in a given situation.

The resulting picture of fragmentation that I've proposed in this paper has, I think, a lot of advantages. Not only is it able to nicely account for all the desiderata that I started out with, but it also very nicely seems to answer some other questions that are tied to fragmentation. For example, my account helps us to tell whether mental fragmentation occurs only in cases in which an agent has incoherent beliefs. According to my proposal this is not the case: For even in situations in which an agent has no incoherent beliefs it might still be the case that she is employing a cognitive process for which some beliefs are easy to access and others are not - and the ones that are difficult to access are in different fragments.

My theory also nicely addresses the question of when we should consider an agent as fragmented. Roughly, we ought to regard an agent as fragmented whenever she employs a particular cognitive process that has a limited domain - such as a specific heuristic or modular process for example. My account leaves it open however that in some cases we do not employ such heuristics or modular processing and so might perhaps *not* be fragmented. Perhaps situations in which we employ what Kahnemann has called System-2 type reasoning might be such situations.[27]

Lastly, my picture also helps us delineate the fragments: In particular, for each situation S the fragments will be delineated by the cognitive process operating in S and the beliefs that are easily (and not easily) accessible to that process.

Additionally, the view I proposed here puts the phenomenon of mental fragmentation in the wider context of cases in which facts about the structure of our cognitive processing excuses (to a certain degree) errors of rationality. Other cases in this group include the cases in which one fails to believe difficult logical consequences of one's view - but there might be others. Notice one important thing about all these kinds of cases: the reason we are less culpable for such mistakes of rationality isn't necessarily because our cognitive processing is limited. For all the cases that I've mention, (and for the cases in which one fails to believe a logical truth or logical consequence of one's beliefs), we can in theory compute these truths and we can entertain all these pieces of information together. It's just that usually we don't - given the way that our cognitive

[27] Though it might also be true that System-2 processes are able to access merely a much larger number of beliefs than System-1 processes, but nevertheless there are many beliefs that are not accessible. In this case we might still be fragmented, but less so (in the sense that the active fragment is much bigger).

I also want to point out one other thing: My account does not imply that an agent who is fragmented is irrational. In fact I want to hold on to the idea that an agent is irrational merely in virtue of violating a requirement of rationality (like the coherence requirement). Fragmentation interacts with questions of rationality only in so far as it might mitigate the irrationality that we attribute to an agent who violates such a requirement. (Though perhaps there is a separate question of whether an agent is irrational in virtue merely of being fragmented - though I don't see why this should be the case.) Thanks for Aliosha Barranco Lopez for helpful discussion on this point.

processing is structured. So fragmentation might also occur in beings that have (in theory) unlimited cognitive computing power - as long as they sometimes use short-cuts (or similar things).

REFERENCES

Carruthers, Peter (2006a). The Architecture of the Mind: Massive Modularity and the Flexibility of Thought. Oxford University Press UK.

Carruthers, Peter (2006b). The case for massively modular models of mind. In Robert J. Stainton (ed.), *Contemporary Debates in Cognitive Science*. Blackwell.

Cherniak, Christopher (1983). Rationality and the structure of memory. *Synthese* 57 (November):163-86.

Davidson, Donald (1982). Paradoxes of Irrationality. In *Problems of Rationality*. Oxford University Press.

Davidson, Donald (1985). Incoherence and irrationality. *Dialectica* 39 (4):345-54.

Dennett, Daniel (1984). Cognitive wheels: The frame problem of AI. In Christopher Hookway (ed.), *Minds, Machines and Evolution*. Cambridge University Press.

Egan, Andy (2008). Seeing and believing: perception, belief formation and the divided mind. *Philosophical Studies* 140 (1):47 - 63

Elga, Adam (2004). On overrating oneself... And knowing it. *Philosophical Studies* 123 (1-2):115-124.

Elga, Adam & Rayo, Agustin (ms a) Fragmentation and information access. (manuscript)

Elga, Adam & Rayo, Agustin (ms b) Fragmented Decision Theory. (manuscript)

Greco, Daniel (2014). Iteration and Fragmentation. *Philosophy and Phenomenological Research* 88 (1):656-673.

Kahneman, Daniel (2013). Thinking fast and slow. Penguin Random House UK

Lewis, David (1982). Logic for equivocators. *Noûs* 16 (3):431-441

Norby, Aaron (2014). Against Fragmentation. *Thought: A Journal of Philosophy* 3 (1): 30-38.

Solso, Robert & Maclin, Otto & Maclin, M. Kimberly (2008). Pearson Education Inc (8th edition)

Quilty-Dunn, Jake & Mandelbaum, Eric (forthcoming). Against dispositionalism: belief in cognitive science. *Philosophical Studies*:1-20.

Schwitzgebel, Eric (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91 (4): 531-553.