Seeing and Believing:
Perception, Belief Formation and the Divided Mind[1]

Andy Egan
Australian National University and University of Michigan
Draft of February 13, 2008

Introduction

On many of the idealized models of human cognition and behavior in use by philosophers, agents are represented as having a single corpus of beliefs which (a) is consistent and deductively closed, and (b) guides all of their (rational, deliberate, intentional) actions all the time. In graded-belief frameworks, agents are represented as having a single, coherent distribution of credences, which guides all of their (rational, deliberate, intentional – this qualification hereafter omitted) actions all of the time. (For example, models in which agents' beliefs are represented by a set of possible worlds that the agent takes seriously as candidates for actuality are like this. So are models in which we use a distribution of credence over a space of possible worlds to reflect a graded taking-seriously relation which allows for the possibility that, among the possibilities the agent takes seriously, some are taken more seriously than others.)

It's clear that actual human beings don't live up to this idealization. Actual human beings don't have a single coherent system of beliefs – either binary or graded – that guides all of their behavior all of the time. The systems of belief that we in fact have are *fragmented* or, as Stalnaker (1984) and Lewis (1982) put it, *compartmentalized*. 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

[1] Thanks to Alan Hajek, Martin Davies, John Campbell, and audiences at Victoria University of Wellington, the Australian National University, and the 2007 Pacific APA for helpful discussion, comments, questions, and objections, in particular to Adam Elga for the ongoing series of conversations that has shaped most of my thinking about these questions.

of our behavior in different contexts. Among the indications of this are the facts that actual people have inconsistent beliefs, display failures of closure, and often fail to bring to bear some of the things that they believe on particular decisions – all phenomena that a single-coherent-system account can't model, but a fragmented account can.[2]

It's tempting to think that, while of course people *are* fragmented, it would be better (from the perspective of rationality) if they weren't, and the only reason why our fragmentation is excusable is that we have limited cognitive resources, which prevents us from holding too much information before our minds at a time. Give us enough additional processing capacity, and there'd be no justification for any continued fragmentation.

Here is one way to illustrate the thought: Imagine two machines. Both do some representing, and both do some acting on the basis of their representations. One is a very fast, very powerful machine. It keeps all of its stored information available all of the time, by just having one big, master representation of the world, which it consults in planning all of its behavior. It also updates the single representation all at once, to maintain closure and consistency: as soon as it starts representing P, it stops representing anything incompatible with P, and starts representing all of the consequences of P. It's just got one, always-active, instantly-updated representation of the world.

The other machine is a smaller, slower, less impressive machine. It can't hold a lot of information in its working memory. So it divides up its store of information about the world

[2] There are two different ways to think about these issues, and I would like, as much as possible, to remain neutral between them. The *radical interpretationist* view is that all there is to facts about belief (in both its binary and graded varieties) is facts about idealized interpretation – what it takes to believe that snow is white is for the best interpretation of someone in your situation, with your dispositions to behavior, to attribute a belief that snow is white to you. The *realist* view is that facts about belief outstrip the facts about interpretation. On a realist account, the question of fragmentation is the question of about how the physically realized mechanisms of representation in our heads are configured – whether or not there are distinct belief-systems guiding different aspects of our behavior in different contexts. On a radical interpretationist account, the question is whether the best interpretation of people will attribute to them more than one system of beliefs, different ones of which are effective in guiding different aspects of their behavior in different contexts.

into lots of little representations, and calls up different ones at different times to guide particular bits of its behavior. Not all of its information is available to it all the time, or for every purpose, and so it sometimes fails to act on things that, in some sense or other, it "knows" or "believes". Since its information about the world is divided into many different representations, updates to one representation don't always percolate all the way through the system. Sometimes this machine "learns" that P, but fails to erase all of its old representations of things that are incompatible with P. Sometimes it represents that P (in one representational item), and that if P then Q (in another representational item), but fails to represent that Q, since the two representational items don't communicate with one another in the right sort of update-inducing way.

It's natural to think that the first sort of machine will always perform better than the second, and that, if you've got the computational resources available, it would always be better to build the first sort of single-representation machine. It's natural to think that the *only* reason why it would be a good idea to design a machine with the second, fragmented sort of representational architecture is if you are stuck with hardware that doesn't have the computational resources to implement the first sort of unified representational architecture.

We are, of course, machines of the second kind. And we're stuck with being machines of the second kind, since we're stuck with an acutely finite computational apparatus. Given the limits on our cognitive resources imposed by our small, slow heads, it's good that we're fragmented. But it's natural to think that if we *weren't* stuck with that – if our heads were enough bigger and faster – it would be bad to be fragmented.

I'll argue that this is not so. There are good reasons to be fragmented rather than unified, independent of the limitations on our available processing power. In particular, there are ways our belief-forming mechanisms – including our perceptual systems – could be

constructed that would make it better to be fragmented than to be unified. And there are reasons to think that some of our belief-forming mechanisms really are constructed that way.

1. Fragmented Belief

In this section I'll present some reasons why we ought to take fragmentation to be the norm among actual human believers, and say a bit about how to think about it. I'll discuss the case for fragmentation in a framework with a binary, on/off notion of belief first, in order to keep things as simple as possible, and then move to the degreed notion once we've already got most of the moves on the table.

One kind of case that makes trouble for the idealized, single-coherent-body-of-beliefs view is the sort of case of inconsistent belief that David Lewis discusses in "Logic for Equivocators" (Lewis 1982). One of Lewis's central examples is a resident of Princeton who believes that Nassau St. runs North/South, believes that the railroad tracks run East/West, and believes that Nassau Street and the railroad tracks run parallel to one another. This sort of pattern of inconsistent belief is probably familiar – we catch ourselves in this sort of situation all of the time.

Lewis proposes that the right way to understand what's going on in such cases is that the agent is in two distinct belief-states, one according to which the street and the tracks run parallel and they both run North/South, and another according to which the street and the tracks run parallel and both run East/West. Sometimes, the agent's behavior is guided by the first belief-state, and sometimes it's guided by the other.

If we want to maintain something like the usual sort of link between belief and action, this looks like a much better option than attributing a single, inconsistent system of beliefs to our agent. Pretty much all of the plausible stories about this connection start by saying that,

when people act intentionally, on the basis of their beliefs and desires, they are disposed to act in ways that would bring about their desires if their beliefs were true. We are inclined to attribute inconsistent beliefs to an agent when they act sometimes and in some respects in ways that would be advisable, given their desires, if P, and at other times and in other respects in ways that would be advisable, given their desires, if not-P. What motivates us to attribute inconsistent beliefs to someone is *not* that they act always, and in every respect, in ways that would be advisable, given their desires, if both P and not-P. We don't, typically, even have a grip on what this last sort of behavior would be like. Which of the actions available to me are the ones that would (tend to) bring about the satisfaction of my desires if P and not-P? (If Nassau Street and the railroad tracks were parallel, and ran both North/South and East/West?) This sort of question is unlikely to lead to answers that will be of any help to us in attributing beliefs to one another.

Nor will they be of much help in the project of predicting people's behavior based on what we know about their beliefs. The prediction I should make about a resident of Princeton, if I learn that he believes that Nassau St. runs North/South, that the tracks run East/West, and that the tracks and the street run parallel, is that he will sometimes act as if both run North/South, and sometimes act as if both run East/West – not that he will always act as if Princeton were arranged in some geographically impossible way (to the extent that we have any grip on what acting that way would amount to).

The cases that motivate the attribution of inconsistent beliefs – cases of acting in some contexts and in some respects in ways that would be advisable if P, and in other contexts and/or other respects in ways that would be advisable if ¬P – provide one sort of motivation for taking people to have fragmented or compartmentalized beliefs. There are

also motivations that don't involve any sort of apparent *inconsistency*, but only failures to "put the pieces together".

One sort of case is failures of closure. Alfred seems to believe that P, seems to believe that if P then Q, but *doesn't* seem to believe that Q. These sorts of cases are also best handled by treating Alfred as fragmented – some bits of his behavior are guided by a system of beliefs that includes P, while other bits are guided by a system that includes *if P then Q*, but none is guided by a system that includes both.

There is a third sort of case, distinct from inconsistency and failures of closure, that motivates a fragmented account of human belief. These are cases of beliefs that seem sometimes, but not always, to be effective in guiding an agent's behavior. The common-sense, folk-psychological distinctions between *recognition* and *recall*, and between *ignorance* and *failure to bring to bear*, rely on the possibility of fragmentation, since they rely on the possibility of different subsets of our total corpus of beliefs being effective in guiding different aspects of our behavior at different times.

Asked, "what was Val Kilmer's character's callsign in *Top Gun*?", I draw a blank. Asked instead, "was 'Iceman' Val Kilmer's character's callsign in *Top Gun*?" I confidently reply, "yes, of course". Why didn't I just say "Iceman" in response to the first question? Not because was ignorant of the fact that – i.e., failed to believe that – Val Kilmer's character's callsign was "Iceman". Of course I believed that. That's why I said "yes" to the second question. It's not as if I learned some new fact, of which I had previously been ignorant, when asked the second question.[3] The reason why I drew a blank when the question was formulated in the first way was that the relevant belief wasn't being brought to bear, not because it wasn't present. That is: It wasn't that I failed to be in any belief-state that included

[3] Of course there are some cases like this – cases in which the way a question is asked allows one to figure out what the correct answer is. But not *every* case is like this.

the proposition, *that Val Kilmer's character in Top Gun's callsign was 'Iceman'*. It was that the belief-state that *was currently guiding my behavior* when I was trying to answer the question didn't include that proposition.

A natural way to think about this: The information was filed somewhere in my head all the time, but it wasn't "called up" – that is, it wasn't available to help guide my question-answering behavior – when I was asked the first question. It *was* called up – it was made available to help guide my question-answering behavior – when I was asked the second question. (The fact that it's difficult to come up with correct answers in the "six degrees of Kevin Bacon" game, but easy to recognize correct answers when they're presented, is another instance of the same phenomenon. We see the same asymmetry – in which it's easier to recognize a correct answer than to generate one – in crossword puzzles, trivial pursuit questions, etc.)

The moral to draw from all this is that our beliefs aren't happily modeled with a single consistent set of believed propositions, or a single set of taken-seriously worlds. A better alternative is to say that we haven't just got a single corpus of beliefs. Our beliefs are *fragmented*, meaning that we've got, each of us, more than one corpus of beliefs, and different ones run different bits of our behavioral show in different contexts. One way to think about this is in terms of the image of calling up files alluded to above: We each have in our heads (or it's as if we each had in our heads) a bunch of distinct representational items, each storing some part of all of our beliefs about the world. Only some proper subset of these representational items is accessed at any given time, by any given behavior-guiding mechanism. And so my behavior-guiding mechanisms only have access to some fragment of the total body of information that I've got in my total system of mental files at a given time.

The initially appealing way to characterize (or at least, begin to characterize) the behavior-guiding role of belief and desire was: Agents are disposed to act in ways that would satisfy their desires if their beliefs were true. (People who believe that P and desire that Q are disposed to act in ways that would bring it about that Q if P were true.) In a fragmented context, this will be replaced with something like: Agents are disposed to act in ways that would satisfy their currently-active desires if their currently-active beliefs were true. (Better: Agents are disposed to act, in a context c and within a domain d, in ways that would satisfy their <c,d> active desires if their <c,d> active beliefs were true.)

(It should be clear from this way of setting things up that I think we'll want to allow for fragmentation of desire as well as belief. I won't, however, say anything else about fragmented desire here. I think it's an interesting and under-discussed phenomenon, but any substantive discussion of it would take us too far afield.)

There will be a corresponding revision to the (beginning of) the initially attractive story about what it is to be a believer: Rather than saying that x believes that P iff x is in some state that represents that P and is disposed to cause x to act in ways that would be successful if P, we should say that x believes that P iff x is in some state that represents that P and is disposed, when active in behavior-guidance, to cause x to act in ways that would be successful if P. (See for example Stalnaker 1984.)

On this sort of picture, we can still characterize our beliefs and desires in terms of sets of worlds,[4] or in terms of consistent, closed sets of propositions. But now we'll need more than one such set of worlds or propositions per person, and we'll need to attach something to each set of worlds or propositions that specifies the scope of its behavior-guiding role – something that specifies the circumstances and domains of behavior in which it's active. We

[4] Officially, I think we'll need to use centered worlds rather than worlds, but since those issues are orthogonal to these, I'll ignore this complication.

need to specify the scope of the behavior-guiding roles for our various belief-fragments in order to fully characterize our overall doxastic states. Just specifying the contents of the various fragments, without specifying the scopes of their behavior-guiding roles, won't tell us anything about what our subject is likely to do.

We'll need to add some further complications in order to tell the right story about revision – about which sorts of inputs the various fragments are sensitive to, and in what way. We'll also need a story about the process(es) by which distinct fragments can be *integrated*. This will, no doubt, all wind up getting pretty complicated. For now, though, let's ignore integration and think about revision and updating this way: fragments have associated with them not just a specification of the scope of their behavior-guiding role, but also a specification of the belief-updating mechanisms whose deliverances they're sensitive to. (We may also need a specification of the contexts in which the different updating mechanisms are effective – a given fragment might be sensitive to updating by *this* system in *these* circumstances but not in *those*.)

A natural normative moral to draw from the evidence that we are, in fact, fragmented, is the one I'm concerned to argue against: That we have a rational failing, but it's excusable because (and only because) of our limitations in processing power. We don't have the processing power to manipulate informationally super-rich representations all the time, and so we can't always keep in mind everything that we believe. Given that, we can be let off the hook for failing to have a single, consistent corpus of beliefs that governs all of our behavior. But if we didn't have the sharp cognitive limitations that we do – if we had the processing power to manipulate representations that encoded all of our beliefs quickly and easily enough – then we'd have no excuse, and fragmentation would be something that we'd be well advised (and rationally obliged) to get rid of if we could.

The discussion of fragmentation above has been in terms of a binary notion of belief. But belief isn't really just an on/off matter. What changes when we move to a more sophisticated, graded notion of belief? For our purposes, not much. The standard sort of idealization (in, e.g., Bayesian formal epistemology and decision theory) attributes to agents a single coherent credence function. The same phenomena that spoke in favor of a fragmented picture of binary belief also speak in favor of a fragmented picture of graded belief, by suggesting pretty strongly that actual people don't approximate this ideal terribly closely, and that there are big chunks of our behavior that fit badly with this idealization, and can't be happily modeled by a theory that retains it.

The required modification of the theory is very similar, as well: We still model agents' beliefs with credence functions, but now an agent's doxastic state will be represented not with a single credence function, but with a set of pairs of credence functions and items that specify the scope of that credence function's behavior-guiding role.

(We can think of each credence function as capturing the way some particular representational item, or collection of representational items, represents (probabilistically) things as being, and the specification of the scope as specifying the circumstances under which various of the agent's behavior-guiding mechanisms "call up" that representational item, or that collection of representational items.)

Cases of inconsistent binary belief become cases in which an agent is quite confident that P, and quite confident that ¬P, and so $c(P)+c(\neg P)>1$. Somewhat more complicated violations of the probability axioms would be required were we to attribute a single credence function to the agent in cases of closure failure.

The sorts of recognition/recall, and failure-to-bring-to-bear cases that we looked at before are also troublemaking cases here. When I draw a blank in response to "what was Val

Kilmer's character's callsign?" and respond with a confident "yes" to "was Val Kilmer's character's callsign 'Iceman'?", what credence, exactly, should we say that I assign to the proposition *that Val Kilmer's character's callsign was 'Iceman'*? There seems to be no happy answer to give – I'm disposed to act in some circumstances and in some respects like someone with a very high credence, and in other circumstances and other respects like someone with a much lower credence.

A unified picture has trouble accommodating this. A fragmented picture has something very natural to say: Much as before, we capture the idea that the information– always filed away in my head somewhere – is successfully "called up" in response to the one question, but not the other, by allowing my verbal behavior in the two cases to be guided by different credence functions, one of which includes a quite high credence in Val Kilmer's character's callsign being "Iceman", and one of which does not.[5]

Finally, in the degreed-belief context as well, we'll want to specify, not just the contexts and domains of behavior with respect to which a given credence function is active in guiding the agent's behavior, but which mechanisms of belief formation and updating a given credence function is sensitive to.

The moral, and the response, are familiar: People's beliefs aren't happily modeled with single coherent credence function. If we want to capture the recalcitrant phenomena, we need to revise the idealized theory, and say that we've got more than one coherent credence function. So we tell the usual decision-theoretic story about how beliefs and desires conspire to produce behavior, except with different credence functions (and, I expect, value functions) running the show at different times and in different domains. A natural first pass at this story

[5] This is a close relative of the sort of picture advocated by Adam Elga (forthcoming).

is the one above: Represent people's doxastic states not just with credence functions, but with a set of <credence function, behavior-guiding scope> pairs.

Why say that rather than assign a single, incoherent credence function? There are two problems: First, this sort of proposal seems to face the same difficulty in generating any predictions about behavior (or prescriptions about what sort of behavior would warrant the ascription) as the proposal that we attribute to our Princetonian the (binary) belief that the streets and railroad tracks are arranged in some geometrically impossible way.

A second problem with this idea is that it doesn't let us capture patterns where the agent reliably behaves, in *this* sort of context or in *this* domain of behavior, like a confident P-believer, but in *that* context or *that* domain, like a confident not-P believer. Assigning incoherent credences either doesn't predict anything or predicts a mish-mash of incoherent behavior when the truth or falsity of P is relevant to the agent's behavior. In particular, it predicts the *same* mish-mash across all contexts, and in all aspects of the agent's behavior. And this won't allow us to capture the phenomena that we actually find. Assigning several coherent credence functions predicts acting, at certain times and in certain respects, like somebody with *these* kinds of opinions, and at other times and in other respects like someone with *those* kinds. The second seems more like what we actually see.

Put another way: What would somebody have to do in order to warrant attributing to them a single incoherent credence function? What would they have to do, for example, to warrant attribution of $c(P)=.9$ and $c(\neg P)=.9$? I am not optimistic about the prospects for answering this question. What would somebody have to do to warrant attributing to them more than one credence function, each with its designated scope of behavior-guidance? *This* question I can see how to answer: they would have to act, in some contexts and in some respects, in the ways that would warrant a flat-out assignment of high credence to P, and in

other contexts and/or other respects in the ways that would warrant a flat-out assignment of high credence to ¬P. And this answer seems to be one that fits nicely with the behavior we see from ourselves and our fellow human beings.

Why several coherent credence functions rather than a single coherent credence function that changes a lot? First, because this forces us to attribute *extremely* bizarre and irrational policies of belief updating to ourselves and our neighbors, and it doesn't allow us to give a very happy account of the phenomena that motivated the move away from a unified, rationally updated credence function, such as the distinction between ignorance and failure to bring to bear. What we see in actual believers looks more like (and is most charitably interpreted as) a pattern of more-or-less rational updating of fragments, in which not every update effects every fragment, than a pattern of crazily irrational updating of a single corpus of beliefs. Second, because we see what looks like *synchronic* fragmentation, in different domains of behavior. (For example, in Freudian cases where the agent's sincere testimony seems to be responsive to one system of beliefs, while their non-verbal behavior is responsive to another.)

So we ought to say, in a degreed-belief as well as a binary-belief context, that our belief-systems are fragmented. And in the degreed-belief context as well, it's attractive to say that this sort of fragmentation is a rational failing that's only excused by our cognitive limitations. Credence functions are, somehow or other, products of configurations of our representational apparatus – of ways of deploying our internal representations that encode the information that we have about the world. We'd expect to see a single credence function governing all of our behavior only if all of our information was instantly available to us all the time – if there was just a single, massive representation encoding everything we know, believe, or suspect, which was always online, governing our behavior. But it would be crazy

to expect our representational architecture to be like that. Not all of our information is called up all the time. So, we should expect to see (expect people's behavior to warrant attributing) several different credence functions governing different bits of people's behavior in different contexts, depending on which subset of all the information that they we have is currently active.

And again, a version of our target thesis looks very plausible: It looks very plausible that, if we had bigger, faster brains, it would be better to just have the single representation, encoding all our information, guiding all of our behavior all of the time, and receptive to every update. The only reason why it's okay for us to fragment is because our heads are small and slow.

This attractive thought isn't right. Fragmentation is a good response to the presence (or possible presence), of certain kinds of imperfectly reliable belief-forming mechanisms.

2. Cartesian and Spinozan Belief-Forming Mechanisms

In this section I'll introduce a distinction between two kinds of belief forming mechanism – *Cartesian* mechanisms and *Spinozan* mechanisms – due to Daniel Gilbert (Gilbert 1991, Gilbert et al. 1993). Later, I'll be arguing that fragmentation is a good response to the presence of less than perfectly reliable Spinozan belief-forming mechanisms. Again, I'll start off by describing the distinction within a binary belief framework.

The most natural way to think of the operation of our belief-forming mechanisms (such as perception) is probably in accordance with what Gilbert (1991) calls the *Cartesian* model: A belief-forming mechanism – for example, visual perception – presents us with a candidate object of belief. Suppose I have a perceptual experience with content P. I then entertain, without yet *believing*, the content of the perceptual experience, and either opt to

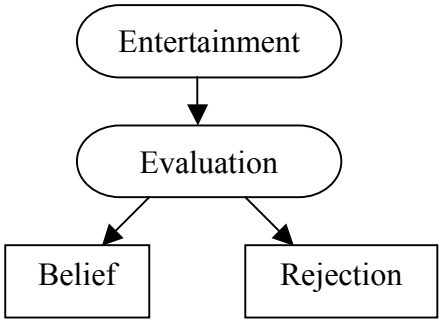
take it at face value, in which case I come to believe that P, or else I reject it, in which case I don't come to believe that P. This is a model according to which adding the deliverances of the system in question to belief is an extra step – once the system's made its deliverances, further processing is required for them to be approved as beliefs.

An alternative possibility is what Gilbert calls the *Spinozan* model: The deliverances of a Spinozan belief-forming mechanism are not subject to any process of evaluation prior to being believed. The mechanism's outputs are directly, automatically, and non-short-circuitably added to the subject's stock of beliefs, though they may subsequently be rejected if they are later subjected to unfavorable evaluation. This is a model according to which belief is automatic, and *rejection* requires a further step, and additional processing.

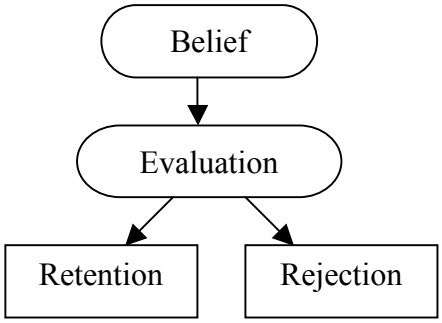
Gilbert uses a "tagging" metaphor to illustrate the distinction. Think of visual perception as producing representations with a certain content. Representations can be tagged in different ways, and play different roles in our cognitive economies depending on which tags they bear. For example, they start playing a belief-ish role in the guidance of behavior once they're given a *belief* tag. One possibility is that the representations produced by the visual system emerge without any tag, and need to undergo some process of evaluation before they get tagged as beliefs. The default state, in other words, is non-belief, and it requires a further process (which presumably will require some cognitive resources) for the deliverances of perception to pass into belief. This is the Cartesian model. Another possibility is that they emerge already tagged as beliefs. The default state is belief, and it requires a further process (which presumably will require some cognitive resources) to *remove* the deliverances of visual perception from belief. This is the Spinozan model.

The difference can be illustrated by the following diagrams (these more-or-less duplicate diagrams in Gilbert 1991):

Cartesian:



Spinozan:



Gilbert only talks about binary, on-off belief. But it's pretty straightforward to apply his distinction to graded belief. We just need to have more options available as possible results of the evaluation stage – rather than just allowing *belief* or *rejection*, we allow the evaluation stage to determine just what probability to assign to the deliverances of the mechanism. (The most natural way to implement this will probably be in terms of Jeffrey conditionalization.) In the graded-belief case, there's the possibility of the evaluation mechanism not just making a decision about *whether* to give credence to the deliverances of the mechanism, but also a decision about *how much* credence to give it. On the tagging model, moving to a graded notion of belief just requires a more complicated system of tags. In a Cartesian system, the deliverances of the system are untagged until evaluated. In a Spinozan system, the deliverances of the system are automatically given some quite high-probability tag, which can subsequently be removed or replaced by a lower probability tag if they're later subjected to unfavorable evaluation.

Gilbert's experiments (1993) give us some reason to think that some of our actual belief-forming mechanisms are Spinozan. The idea behind the experiments is to test to see what the default, pre-evaluation outcome of the operation of the mechanism is, and what requires an additional, cognitive-resource-demanding evaluative process, by seeing what

happens when the mechanism operates while the subject is under a cognitive load. The idea is that this will make it more difficult to go through the *evaluation* stage, and so we'll see, disproportionately, the default output of the mechanism rather than its post-evaluation output. If the mechanism is Cartesian, placing the subject under cognitive load should make it harder for the deliverances of the mechanism to make it into belief, by preventing them from undergoing evaluation. If the mechanism is Spinozan, we should see *more* of the mechanism's deliverances retained as beliefs, since some of the deliverances which would have been rejected under evaluation will slip through when the process of evaluation is hindered by a lack of available cognitive resources.

Gilbert's experiments suggest that reading comprehension, for example, is Spinozan. In one experiment (see Gilbert et al. 1993), subjects are instructed to read a passage, and told that the sentences in the passage printed in red are false, while those printed in black are true. Some subjects are asked to perform an additional, cognitively demanding task while reading the passage, while others are allowed to read the passage uninterrupted. The subjects who are not performing the additional task wind up believing (pretty much) only the bits of the passage printed in the reliability-indicating color. Those who are reading while performing the additional task are almost equally good at taking on board the truth-marked claims, but much worse at filtering out the falsity-marked ones – they are much more likely to wind up believing the claims that were printed in the falsity-indicating color. This suggests that *rejecting* things we read takes more work than accepting them. The presence of the extra task should interfere with any additional processing that's required to give the deliverances of reading comprehension something other than its default status (to change its tag). The findings fit well with an account according to which the default status is belief. None of this is conclusive, but it's at least suggestive.

If any of our belief-forming mechanisms are Spinozan, the various sorts of perception are likely to be among them. For one thing, they often deal in information that it would be a good idea to take on board *right away*, with a minimum of fuss. When visual perception reports that a tiger is charging out of the underbrush, one does better if one doesn't pause (even for a moment) to subject its deliverances to evaluation before turning to run. (Or devote cognitive resources that could be used for planning one's escape to deliberating about whether or not to take the tigerish visual appearance at face value.) Better to believe it straightaway and run (climb a tree, ready a weapon…) immediately. So it seems as if it's at least a live possibility that at least some sorts of perception are Spinozan, and seeing really *is* believing.

In a fragmented context, we need to distinguish between *selective* and *broad-spectrum* belief-forming mechanisms. A *selective* belief-forming mechanism acts only on some proper subset of the subject's belief-fragments or belief-systems. A *broad-spectrum* belief-forming mechanism acts on all of the fragments at once. We're not likely to be interestingly fragmented if all of our updating mechanisms are broad-spectrum. To be fragmented in interesting ways, you pretty much need to have belief-forming mechanisms that act *selectively* on your different fragmented systems of belief. So even if a given belief-forming mechanism invariably produces a belief that P (gives one a credence-function that assigns high probability to P), it's pretty unlikely that it will invariably make it the case that *all* of your fragments include P, or that all of your fragmented credence functions assign high probability to P. And there's an interesting distinction, even for belief-forming mechanisms that only act selectively on some of one's fragments, between the ones that do so in an automatic, non-short-circuitable way, and the ones that do so only after their deliverances are

subject to some sort of evaluation. In a fragmented context, belief-forming mechanisms – whether Spinozan or Cartesian - are unlikely to be broad-spectrum.[6]

We could have a non-short-circuitable belief-forming mechanism that forces high credence in its deliverances. This could even be a good thing to have, given our limited cognitive resources. If it always took extra cognitive effort to get the deliverances of perception into belief, we'd be slower and less reliable in our uptake of perceptually given information, especially when some of our cognitive resources are occupied by doing or thinking about other things. It is plausible that, in many cases, there are higher costs to failure, or delay, of uptake than to the occasional uptake of falsehoods. And, given general facts about the fallibility of the sorts of belief-forming mechanisms we've got access to, we could have such a mechanism that wasn't completely reliable, or one that was systematically unreliable in certain kinds of contexts.

3. Coherence and Self-Trust

There are sharp limits on a unified, consistent believer's capacity to mistrust her own belief-forming mechanisms. Obviously, no consistent system of (binary) beliefs contains both *P* and *not-P*. Also, none contains all of *P*, *my belief that P was formed by method M*, and *every belief formed by method M is false*. In fact, consistent systems of binary belief don't seem to allow for *any* epistemic self-doubt. Famously, there is no consistent set of beliefs that contains each of the claims made in some very long book, and also the proposition that at least one of the claims in the book is false. This isn't a special fact about books. It's a general fact about potential sources of information – it's not possible for a

[6] There's also the complicating (and hereafter ignored) possibility of mixed belief-forming mechanisms, whose influence *here* is automatic, but whose influence *there* is contingent on positive evaluation.

unified, consistent agent to both believe all of the deliverances of some source of information *and* believe that the source is less than perfectly reliable.

The situation in the case of graded belief is similar, though not as extreme. There are sharp limits on the credence that an agent who is unified and coherent, and whose credences are transparent to her, can assign to hypotheses about the unreliability of her own belief-forming mechanisms while still assigning high credence to the deliverances of those mechanisms. (See Egan and Elga 2005.)

Consider for example hypotheses of *anti-expertise*. Let us characterize a *subject matter* as set of propositions, and let us say that an agent has a *confidently held belief* that P iff they have a credence function c such that $c(P) > .9$. Now let us say that one is an *anti-expert* about a subject matter S iff at least half of one's confidently held beliefs about S are false. I cannot (a) be well-informed about my beliefs, (b) have a confidently-held S-belief, and (c) assign $c > .2$ to hypothesis that I'm anti-expert about S.

This is an instance of a general theorem (for details, see Egan and Elga 2005), but the guiding idea can be illustrated in the special case of a subject matter that consists of exactly two propositions. Suppose that an agent has 90% confidence in each of the two propositions. Then she must be at least 80% confident in their conjunction (for the same reason that two carpets, each of which covers 90% of the floor space of a single room, must overlap on at least 80% of the space). So she must be at least 80% confident that she is correct in both of her confident beliefs on the subject matter—and hence can be at most 20% confident that she is an anti-expert.

Given this constraint, if an agent who is always unified and coherent, and whose beliefs are transparent to her, starts off with a confidently held belief about a subject matter S (call this an *S-belief*), she cannot become convinced that she is an anti-expert about S without

giving up her S-belief. If she starts off convinced of her anti-expertise, she cannot acquire a confidently held S-belief without reducing her credence in the hypothesis of anti-expertise. In transparent (even reasonably transparent) agents, confident S-beliefs and significant credence in anti-expertise about S can't coexist – adopting the one will force them to give up the other.

4. When it's Good to be Fragmented

Let's start with a cartoon example: Bill is a unified, coherent agent with a Spinozan visual system. One day, Bill gets extremely compelling evidence that his visual system is very unreliable in a certain sort of circumstance – in the kitchen, say. As a result of taking this evidence on board, he becomes convinced that his visual system is unreliable in the kitchen. He then wanders into the kitchen and opens his eyes, and his visual system makes some deliverances.

What's going to happen to Bill? Since his visual system is Spinozan, he can't help but believe its deliverances (i.e., he can't help but assign them high credence). So, Bill straightaway becomes confident that the kitchen is full of fruit bats. Since Bill is coherent and unified, his confident belief in his visual system's unreliability can't coexist with his confident belief of the system's deliverances, he's forced to discard his belief that the system is unreliable. (In a graded belief framework, he's forced to give it quite low credence.)

He's also forced to reduce his credence in anything that would be sufficiently good *evidence* that his visual system was unreliable in the kitchen. (Since where c(P|Q) is sufficiently high, imposing a low upper bound on c(P) will also impose a low upper bound on c(Q).) So, for example, his moving to very high credence on the kitchen's being full of fruit bats will force him to have very low credence that the studies establishing the unreliability of

visual perception in the kitchen really were conducted carefully by reputable scientists, etc. He's likely to come to believe, not just that the room is full of fruit bats, but that the scientists who conducted the studies he trusted a moment before were charlatans. (Or that his apparent memories of reading the kitchen-vision discrediting papers aren't veridical, or that *his* visual system is very special, or...)

In general, if you're a unified believer with an unreliable Spinozan belief forming mechanism, no amount of pre-exposure education about the unreliability of mechanism is going to be of any help to you once the mechanism starts making its deliverances. As soon as the mechanism's deliverances come in, they'll crowd out all of your hard-won confidence of the mechanism's unreliability, and all of your confidence in the evidence that convinced you of its unreliability in the first place. Neither your belief that the mechanism is unreliable, nor the collection of supporting beliefs about the evidence that convinced you to believe that the mechanism is unreliable, is going to survive any contact with the enemy. Forewarning of the unreliability of their Spinozan belief-forming mechanisms leaves the agent who is always unified and coherent just as badly armed as they were before.

That's bad. If you're at risk of having an unreliable (or situationally unreliable) Spinozan belief-forming mechanism, it's good to be able to find out about it. And it's good if your knowledge that the mechanism is (situationally) unreliable can actually do you some good in helping you to discount its deliverances when they come in, rather than being immediately washed out once the mechanism starts making its unreliable deliverances.

You'll do okay, though – or at least, you'll be likely to be able to do *better* – if you're able to fragment, so that you're left with one belief-system that contains (high credence in) the deliverances of perception, and another that contains (high credence in) the proposition that the perceptual mechanism is unreliable, as well as the discrediting evidence. Just how

much better you do will depend on which fragments have which functional roles in guiding behavior, and in how likely it is that you'll eventually subject the fragment containing the deliverances of perception to a process of evaluation where it won't pass muster. But at least if you fragment, you won't be stuck with the unreliably-formed belief guiding *all* of your behavior, and you won't have lost all traces of the perception-discrediting beliefs that any later process of evaluation would need to appeal to in order to conclude that the perceptual mechanism was delivering bad information.

So this is a sort of situation – pretty plausibly, our situation – in which it's better to be fragmented than not, completely independent of any issues about our capacities to rapidly process lots of information. Being fragmented, it turns out, is a good idea if you're going to have (or if you're stuck with) Spinozan belief-forming mechanisms that aren't perfectly reliable.

In fact, the *evaluation* stage of the Spinozan model actually only makes sense on fragmented picture. The whole idea of evaluating, based on fit with the rest of your beliefs, whether to retain something as a belief, requires that you be able to represent both the candidate belief and the other – potentially inconsistent – beliefs that it's being evaluated for fit with at the same time. In fact, the whole idea of testing beliefs for fit with one another, and discarding some in virtue of their inconsistency with the rest, relies on a fragmented picture of belief. We can't ever find, on evaluation or reflection, that some of our beliefs are inconsistent with the rest unless it's possible for us to *have* inconsistent beliefs. So this familiar picture of the evaluation of beliefs will only find useful application for fragmented agents.[7]

[7] Thanks to Martin Davies for discussion here.

4. Conclusion

Agents with Spinozan belief-forming mechanisms that might turn out to be less than completely reliable would do well to be fragmented. But it's not only Spinozan systems that generate this kind of trouble. An agent with only Cartesian systems will also do well to be fragmented if their mechanisms of *evaluation* might turn out to be faulty. The problem to which fragmentation is a good response is the problem of having the deliverances of some unreliable system fed directly into a system of beliefs that then loses the ability to criticize and discard the faulty input. It's just as bad to be unable to use one's knowledge that one's *evaluative* mechanisms are unreliable to shield oneself from their bad effects as it is to be unable to use one's knowledge that one's *perceptual* mechanisms are unreliable.[8]

How much being fragmented is likely to help is going to depend a lot on the nature of the fragmentation. Suppose Bill, our subject with the unreliable-in-the-kitchen Spinozan visual system, is fragmented rather than unified. Since his visual system is Spinozan, he's going to form a belief-type representation of the false content of his visual experience. Call that content P. As long as the representation-fragment that contains Bill's evidence that his visual system is unreliable isn't immediately sensitive to updating by his visual system, he'll also retain his belief-type representation of the perception-undermining evidence. (Call this E.) How much good this will do for Bill will depend on the details of the roles of that Bill's P-representation and his E-representation play in his cognitive economy.

One place where this will show up is in the *behavior-guiding* role of the two representations. The more of Bill's behavior is guided by his E-representation, and the less of it is guided by his P-representation, the better things are likely to go for him. (It would be particularly good if Bill's belief in the unreliability of these particular perceptual inputs were

[8] Thanks to Stuart Brock for raising this point.

able to cause the belief-type representations that were the inevitable output of his visual system to have a more restricted behavior-guiding role than perceptually-produced beliefs typically have.)

Another place where the details will matter is in the *updating* role of the two representations – in the way that Bill's cognitive systems address conflicts between the two sorts of belief-type representations. In order for Bill's E-representation to help mitigate the effects of having the P-representation around, it's got to be that the presence of the E-representation has some tendency to bring it about that the P-representation is discarded, revised, or relegated to a less-prominent role in the guidance of behavior.

The phenomenon being pointed out here is that, while an absolutely ideal believer will both (a) have completely reliable mechanisms of belief-formation, and (b) incorporate all of their reliably-obtained information about the world into a single unified corpus of beliefs that's active in guiding all of their behavior all of the time, we shouldn't think that it's a good idea to have the second feature in the absence of the first. Given that we're going to fall away from the ideal with respect to (a), it's *better* for us if we fall away with respect to (b), as well.

This mirrors a general result that's familiar from economics (see e.g. Lipsey and Lancaster 1956). If we're constrained in such a way that the solutions available to us lack one of the features of the best solution, it's not in general true that the second-best solution – the best solution available to us, given the constraint – will retain the *rest* of the relevant features of the best solution. (In fact, it's in general true that the second-best solution will *not* retain the rest of the relevant features of the best solution.) A standard example in economics is that it is 'efficient' to tax all goods at identical rates if you can tax *all* goods (including leisure); but if one good cannot be taxed (leisure say) then it will not be best – i.e., will not be

efficient – to tax all remaining goods at equal rates. (Thanks to Geoff Brennan for pointing this out, and for the example.)

We can see similar phenomena in a wide variety of places. Cases of taking account of our own weakness of will are probably like this. The best thing for Jackson and Pargetter's (1986) Professor Procrastinate to do is to accept the invitation to referee the paper for the journal, and to write a report in a timely manner. But if he knows that he will not, in fact, write the report in a timely manner if he accepts, the best thing for him to do is not accept the invitation. In the political sphere, failures of well-intentionedness are happily accompanied by failures of power. And as Kant points out, the value of such virtues as courage, wit, and persistence is contingent on their being accompanied by a good will. There is also a parallel with group inquiry, in which it's good, if there's a danger of people becoming convinced of false views, for the spread of ideas through the scientific community not to be too rapid or complete. If everybody gets swept up by the fad, the community will lose its capacity for self-criticism.

References:

Egan, A. and Elga, A. (2005) I Can't Believe I'm Stupid, *Philosophical Perspectives* 19, 77-94.

Elga, A. (forthcoming) How to Be Incoherent, and Why.

Gilbert, D. (1991) How Mental Systems Believe. *American Psychologist* 46, 101-119.

Gilbert, D., Tafarodi, R., and Malone, P. (1993) You Can't Not Believe Everything You Read. *Journal of Personality and Social Psychology* 65, 221-233.

Jackson, F. and Pargetter, R. (1986) Oughts, Options and Actualism. *The Philosophical Review* 95:2, 233-255.

Kant, I. (XXXX/XXXX) *Groundwork for the Metaphysics of Morals*

Lewis, D. (1982) Logic for Equivocators. *Nous* 16, 431-441. Reprinted in D. Lewis, *Papers in Philosophical Logic*, Oxford, Oxford University Press, pp. 97-110.

Lipsey, R. and Lancaster, K. (1956) The General Theory of Second Best. *The Review of Economic Studies* 24:1, 11-32.

Stalnaker, R. (1984) *Inquiry*. MIT Press, Cambridge, Mass.