3

The Fragmentation of Belief

Joseph Bendaña and Eric Mandelbaum

1. How Are Beliefs Stored?

Belief storage is often described with two metaphors. The first metaphor is that of a *belief box*, a functionally individuated warehouse where mental representations reside. As functionalists have it, a believer instantiates the belief relation, whatever exactly that relation is, to the propositions encoded by those mental representations.[1] The second metaphor is that of a *web of belief*, according to which all of an agent's beliefs are interconnected in a single, web-like network, the entirety of which synchronically guides action and reasoning. Taking this metaphor at face value, even beliefs that seem deeply unrelated, say one's belief that $9 \times 3 = 27$ and one's belief that one's parents are immigrants, would count as connected to each other in some arcane though consistent way. This metaphor is attractive because it promises to account for how people can reason about many different subject matters at one time (thus allowing for the 'isotropy of belief'; Fodor 1983). If an agent's set of beliefs is encoded in a single web-like data structure, evidence that affects the status of one belief can have consequences for the overall topology of the web, affecting the entire web at once (thus allowing for the supposedly 'Quinean' nature of belief; Fodor 1983). In this way, our beliefs about pop music could, in principle, constrain our beliefs about paleontology (you could have based all of your pop music and paleontological beliefs on testimony from someone you now consider to be a liar, or you could think that velociraptors had the best taste in classic mandopop).

Most criticisms of functionalist theories of belief storage have focused on the first metaphor, the belief box. For instance, Schwitzgebel (2013) argues that

[1] We are assuming that beliefs are stored mental representations characterized by their functional role. Since we want our picture to be consistent with a wide range of views, we are not building much into belief's functional role. We just assume that a belief's functional role includes being semantically evaluable, as well as being capable of entering into inferential transitions and being caused by perception. The informational states that we call 'beliefs' in this chapter meet such conditions. Nevertheless, if the reader wants to reserve the term 'belief' for only normatively respectable or reflectively endorsed states, feel free to replace our term 'belief' with the term 'information.' As long as one thinks that the human mind processes stored information, in order to have a complete cognitive architecture, one will have to confront questions about the structure of that information storage.

Joseph Bendaña and Eric Mandelbaum, *The Fragmentation of Belief* In: *The Fragmented Mind*. Edited by Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, Oxford University Press.
DOI: 10.1093/oso/9780198850670.003.0004

accepting that there is a belief box presupposes a binary notion of belief, such that any proposition is either fully believed or not, and that this is unacceptable because of the existence of partial beliefs. In a similarly skeptical tone, Dennett (1991) argues that talk of a belief box reifies beliefs into concrete particulars, whereas he takes beliefs to be just explanatory abstracta. Churchland (1981) claims that the whole enterprise of providing a functional individuation of propositional attitudes is unscientific. We find these worries to be a bit overblown—the belief box metaphor is not meant to limn the structure of storage or the nature of beliefs (Quilty-Dunn and Mandelbaum 2018a). Instead, it is just a convenient way of noting that propositional attitudes are individuated by their functional roles.

In contrast to the belief box metaphor, the web of belief metaphor—the idea that belief storage has the structure of a single web—remains largely unchallenged. In what follows, we set aside questions about the belief box and instead focus our criticisms on the idea that belief storage is best modeled as a single web. When one scrutinizes human cognition, one finds evidence against a single web of belief and for a fragmented model of belief storage.

2. The Web and Its Discontents

Let the Web of Belief (henceforth: the Web) be understood as a conjunction of the following theses:

(1) Unity: Beliefs are stored in a single database.[2]
(2) Quineanism: Belief revision is sensitive to global properties of an agent's total set of beliefs (that is, the beliefs taken as an entire set).
(3) Consistency: When any of an agent's beliefs change, all other beliefs adjust to remain consistent with the modification.[3]
(4) Uniqueness: Belief storage does not contain redundant token representations.
(5) Conservatism: The more revisions a belief change would require, the less likely it is that the change will occur.

[2] We are assuming that databases are functionally individuated at least in part by the patterns of access they enable. So, from our perspective there isn't an important difference between having two different databases and only being able to access one at a time and having one database but only being able to access half the information in it because of some limitation of the mechanism of access. There is an interesting question here: What kind of evidence could determine whether some limitation in information access was a result of a bad search procedure as opposed to poor information organization? However, addressing this question is a task for another paper, both because of the issue's enormity and because we have absolutely no idea of how to go about addressing it.

[3] We find it easiest to construe consistency primarily as a syntactic notion. If content is more coarse-grained than syntax, the Web would then allow for beliefs with inconsistent content. It would just not allow for beliefs with inconsistent syntax. If the granularity of content matches the granularity of syntax, then consistency of syntax would ensure consistency of content.

Aspects of the Web appear in many discussions of belief.[4] However, we are not very concerned with whether any particular philosopher ever explicitly endorsed all five theses listed above. What is important is just that these theses are implicit in many modern models of cognitive architecture, and they provide us with a clear theory of the structure of belief storage that generates the following predictions:[5]

P1: Because of Consistency, when one acquires new information, all inconsistent beliefs are (*ceteris paribus*) ironed out.[6]

P2: Because Quineanism requires an agent's entire belief set to guide reasoning and behavior, reasoning and recall do not typically exhibit context sensitivity when context shifts are irrelevant to the task at hand.

P3: Because of Conservatism and Consistency, people will be less likely to revise beliefs about logical or mathematical truths than beliefs about nearly anything else. This is because changing logical and mathematical beliefs entails a potentially infinite number of changes to the web.

In opposition to the Web, we offer a fragmented view of belief storage. Let 'Fragmentation' be the conjunction of the following theses:

(1) Isolation: Information utilizable by cognitive processes is stored in distinct, independently accessible data structures, which we will call 'fragments.'
(2) Inconsistency: Any fragment that harbors a belief that P and a belief that $\neg P$ will incite a revision process to eliminate the inconsistency, but having one fragment contain P and a separate fragment contain $\neg P$ need not.[7]

[4] Quine (1951) famously compared the human belief system to a vast interconnected field of forces that is constantly updating in light of experience, and Quine and Ullian (1978) later explicitly described it as a web. How exactly to understand these passages is a contentious matter. According to at least one interpretation, Quine was making a descriptive claim about the structure of human belief storage and the nature of belief revision in line with the Web (see, e.g., Cherniak 1983). Arguably, Fodor (1983) also thought that belief revision and all processes that took place in central cognition required constant access to all stored beliefs. The frequently endorsed generality constraint on concept possession has a similar flavor.

[5] Modern cognitive architectures such as EPIC and SOAR and many Bayesian models of cognition, including some models of categorization, presuppose that if information is encoded in memory, that information is available for guiding task performance (Schultheis et al. 2006; Sanborn et al. 2006; Lewandowsky et al. 2006). As we will see, such models leave many cognitive phenomena to be explained. Our proposal can be seen as a very general framework for how these architectures could be modified to explain data that they currently cannot capture.

[6] 'Ironed out' entails that the competence of the system is such that all explicit, syntactic contradictions should be eliminated (so failing to do so would be a performance error). What it means for contradictions that arise as long chains of reasoning is less clear, though presumably the same performance/competence move is applicable here (and we assume as much in what follows). The subsequent empirical predictions that follow in the main text also allow for the same utilization of the performance/competence distinction.

[7] Consistency is thus maintained within, but not across, fragments. In other words, intrafragment inconsistency detection is automatic, but interfragment inconsistency detection isn't.

(3) Locality: Information updating, including belief revision, normally takes place within a single fragment at a time.
(4) Redundancy: Different tokens of any particular belief may be stored in different fragments.
(5) Multiple Resistance: Beliefs that are most resistant to revision are beliefs that are most redundantly represented.

Fragmentation entails the negation of the Web but also goes beyond mere negation. Isolation entails the denial of Unity, since each fragment is functionally isolated from every other fragment.[8] Because information storage is split up into many different data structures, one can neither update all of one's beliefs at once nor synchronically take all of one's beliefs as an evidential base. Instead, people update a single fragment at a time. Locality thus entails the denial of Quineanism. Redundancy, however, is stronger than the idea that one might have type-distinct concepts that have the same content (e.g. Fodor 1998). Instead, it allows that there could be type-identical but token-distinct beliefs residing in different fragments of the mind, and that the token-distinct beliefs can be processed independently of one another. Finally, Multiple Resistance allows for the possibility that the beliefs that are most resistant to revision are not beliefs about necessary truths—this is important because, as it turns out, the beliefs that humans seem least likely to revise are in fact highly contingent (and often false) beliefs about the self: the belief that one is a good person, a smart person, a reliable, consistent person, and the like.

Fragmentation makes at least the following three predictions, which are directly opposed to the predictions of the Web:

P1: Because of Inconsistency, people will be very likely to store inconsistent information.

P2: Because of Isolation and Locality, information access should be extremely context sensitive, and one should be able to elicit independent responses to the same stimuli merely by changing task-irrelevant aspects of the context.

P3: Because of Multiple Resistance, beliefs that people identify with are apt to be the most resistant to disconfirmation (in the terminology of Railton 2014, beliefs about the self will be the most 'resilient').

Now that we have Fragmentation and the Web on the table, we can move on to the evidence. Below, we argue that the best available evidence from cognitive science bears out the predictions of Fragmentation and thus undermines the Web. We conclude by presenting a positive picture, sketching the functional profile of fragments.

[8] This is a claim about *cognitive* architecture: however the information processing turns out to be implemented in the brain, cognitive processing is only done over subsets of a subject's total stored information.

3. Evidence for Fragmentation

3.1. Redundancy

We begin by focusing on how issues of representational redundancy arise in two disparate areas of cognitive science: reinforcement theory and attitudinal psychology. In both areas, one can observe peculiar patterns of failures to modulate mental representations that are difficult to explain without postulating Redundancy.

3.1.1. Extinction and Reinstatement

Suppose one has been conditioned to associate a conditioned stimulus (CS) (e.g. the ringing of a bell) with an unconditioned stimulus (US) (e.g. a bell). Extinction is the process whereby, if all goes well, one breaks the association between the CS and the US. For all its theoretical prominence, extinction is deeply inefficacious, displaying strange patterns of failures. For instance, change of context can cause an immediate and robust return of a previously 'extinguished' association. This particular kind of context sensitivity is hard to account for with a single web of belief but naturally explained by a fragmented architecture that allows for representational redundancy.

Consider associative renewal. While several versions of this effect have been found, here we focus on the most common—'ABA renewal.' Say a dog forms an association between food and the ringing of a bell in a particular spatial context, *A*. The association is then extinguished in context *B*, typically by repeatedly presenting one of the previously associated stimuli without the other, say the bell without the food. Eventually, the dog will stop salivating at the sound of the bell in context *B*. However, when the dog is returned to context *A*, the bell will once again produce a salivary response. There is nothing special about dogs—the same effect holds for (e.g.) humans. The moral is that a learned association can be destroyed in one context but then rearise in another without any additional learning (Mandelbaum 2015a).

Renewal is robust: it can be observed in almost every type of conditioning paradigm in which it has been investigated (Bouton 2004; Bustamante et al. 2016). It occurs even after extensive extinction training (up to 160 extinction trials; Rauhut et al. 2001). It can be obtained without alterations of spatial context: if enough time is allowed to pass after extinction, the CS–US association can rearise on its own. This phenomenon, 'Spontaneous Recovery,' suggests that alterations in temporal context can have the same effects as alterations in spatial context. Renewal and Spontaneous Recovery (as well as other related effects) can also occur after counterconditioning (Bouton 2004).

Spatiotemporal context thus 'sets the occasion' for the association of a CS and a US (Herrnstein 1969). Different contexts serve as triggers, bringing up certain

associations but not others. For example, someone might activate an association between the CS and US1 in one room but activate an association between the CS and US2 in another. Spatiotemporal context is what determines which information (here, which association)[9] is accessed in any given situation.

Fragmentation offers a natural explanatory framework for these results, as it allows for multiple token representations of CS–US associations housed in many distinct fragments, each activated by different contexts. In essence, Fragmentation sees each two-place associative relation as a three-place relation between the US, the CS, and a context (or between a context, stimuli, and response, depending on one's preferred metaphysics of conditioning). Renewal takes place because returning to the context in which the initial CS–US association was acquired reactivates the fragment that was opened during the initial learning. Since extinction occurred in a different context, this first fragment lies quiescent during extinction and thus remains unchanged by the novel learning. The same sort of explanation holds for Spontaneous Recovery, with recovery driven by changes in temporal context. It is unclear how the Web could accommodate such patterns of information access since it does not permit redundancy and has no other resources for explaining why a change in context would change which information was accessible.

3.1.2. Implicit Bias

Fragmentation can similarly explain seemingly disparate social psychological findings regarding implicit bias. Implicit biases are widely taken to be conditioned associations (though cf. De Houwer 2009, 2019; Mandelbaum 2016). It has thus been a standing mystery why experimenters can modulate a person's implicit bias only to have it recover upon retest after brief delays (Devine et al. 2012; Lai et al. 2016).

Fragmentation can simultaneously explain both the short-term effectiveness and the long-term failure of implicit attitude intervention. Suppose one has a (stereotypically anti-Semitic) association between JEWS and MISERLINESS.[10] Assuming Fragmentation, if that is a strongly held association, one will have many such token associations. However, breaking one of them will only affect the particular fragment that is active during the modification process. Other fragments will still have other associations (or, more realistically, beliefs—see, Mandelbaum 2016; Karlan 2020; Bendaña Chapter 11 in this volume) about Jews—some orthogonal (e.g. that they are athletic), some inconsistent (that they are not cheap), and some mere redundant representations. Alterations in

[9] We aren't especially drawn to the view that what is acquired in conditioning paradigms is an association structure per se, as opposed to a proposition (see De Houwer 2009; Mitchell et al. 2009), but we would prefer to sidestep this question here. The reader should feel free to substitute 'proposition' or 'belief' for 'association.'

[10] Small caps denote concept names.

spatiotemporal context between the first and second test could cause information queries initiated by the different tests (or the same type of test on different occasions; see Vul and Pashler 2008; Payne et al. 2017) to access the information in those yet unmodified fragments. Implicit biases could thus persist even though some of the representations that undergird them have been successfully modified.

3.2. Inconsistency

People have a strange relationship with inconsistency. On the one hand, dissonance theory gained fame by showing the lengths to which people will go to avoid inconsistency. Inconsistency causes literal psychological discomfort, and people's penchant to assuage the pain of inconsistencies is seen as a basic drive (to avoid discomfort; McGregor et al. 1999). On the other hand, people are replete with inconsistent information. Even a paragon of rationality like David Lewis (1982) noted that he held inconsistent beliefs about matters as quotidian as the directions of streets and railroads in Princeton.

Inconsistency is particularly difficult to fit into the Web. Recall that according to the Web, all of the information a person has stored is utilized for action guidance and reasoning at once. Thus, if people believed P and believed $\neg P$, they would have to at one time act and reason as if P and as if $\neg P$.[11] However, the idea of at one time acting and reasoning as if P and $\neg P$ borders on incoherence.[12]

By contrast, Fragmentation allows inconsistent information to be selectively accessed. Recall that Fragmentation requires each fragment to be consistent, without demanding interfragment consistency. As long as inconsistent beliefs are stored in different fragments, they can be stored without having to simultaneously guide action and reasoning. Thus, the more contextually mediated access to inconsistent information we find in human memory and cognition, the more evidence we have for a fragmented model of belief storage.

Psychology is rife with findings demonstrating that people are full of inconsistent beliefs. Thoroughly documenting all such cases would surpass the bounds of this chapter, so we will simply highlight certain underexplored models and results that illustrate some of the stranger patterns of inconsistency that humans exhibit.[13]

[11] Inconsistent beliefs can come in many varieties: one can have a single belief state that has inconsistent content, believing P and $\neg P$; or one can have two beliefs, one of which is the negation of the other: a belief that P and a belief that $\neg P$ (to say nothing of inconsistent triads and the like, which was Lewis's predicament). We are agnostic about the possibility of there being a single belief with inconsistent content. All that is needed for our purposes is inconsistent belief states.

[12] For example, if you believed your bathtub both was and wasn't filled with snakes, how would you manage your hygiene? See also Egan (2008), who wonders what it would even mean to act on inconsistent beliefs.

[13] For the reader interested in other work that highlights inconsistent beliefs, see Lewandowsky and Kirsner (2000), Ripley (2009, 2011), Gweon et al. (2011), Legare et al. (2012), Newby-Clark et al. (2002), Hall et al. (2012), Garcia-Marques et al. (2015), Fazio et al. (2019).

3.2.1. Ballistic Believing

The belief acquisition literature provides evidence that people harbor many inconsistent beliefs. This evidence has been previously discussed at length (Mandelbaum 2014; Mandelbaum and Quilty-Dunn 2015), so here we will simply remind the reader of the main findings. In a series of studies, Dan Gilbert and colleagues (1993) have shown that acquiring beliefs happens automatically, ballistically, and effortlessly, while rejecting propositions is a controlled, effortful, and breakdown-prone process. The belief acquisition literature can be summed up by saying that we end up believing any truth-apt information that we parse, regardless of the modality. That is, we initially (unconsciously) believe every proposition we encounter, and only after can we go back and reject that information if we have the time, attention, and disposition. This 'Spinozan' model of ballistic believing ensures that people will harbor inconsistent beliefs. In everyday life, one cannot help but encounter inconsistent propositions, and since cognitive load is so frequent, no one can be vigilant enough to reject all the propositions that one should reject (Levy and Mandelbaum 2014). Thus, we are bound to regularly acquire many inconsistent beliefs.

Although ballistic believing is the best existing model of the belief acquisition data we have, it is still controversial. Thus, the case for differential access to inconsistent beliefs would be strengthened by other examples. Happily, these examples lend support to the Spinozan model in addition to providing evidence for the claim that humans have contradictory beliefs.

3.2.2. Retraction

A Spinozan architecture is counterintuitive. If propositions are automatically believed, and if cognitive load is an ever-present fact of modern life, then we should be acquiring beliefs without sufficient evidence all the time. Many experimental settings are load-inducing on their own (Gilbert et al. 1990), so experimental participants should believe the information presented to them, even if they know full well that the information is false. Indeed, this happens. Participants in pre-briefing experiments who are told that they are receiving false information cannot help but uptake and utilize that information in inference (Wegner et al. 1985).[14]

Similarly, in cases where information is presented and then retracted, a Spinozan architecture dictates that people will simply accept both pieces of information if cognitive load is present. In a fragmented architecture, retractions given under cognitive load should be acquired in a new fragment, in which case

[14] Pre-briefing experiments are those in which participants are told the purpose of the experiment *before* the experiment, as opposed to a normal experiment, where they are debriefed afterwards. In general, if you are going to lie to participants, you don't tell them that you are going to lie to them and instead debrief them about the lie afterwards (which is why the efficacy of pre-briefing experiments is particularly interesting).

both the retraction and the original belief should persist. By contrast, according to the Web, when a retraction is believed the information retracted should be stricken from our beliefs, and the entire topology of the web should be updated accordingly.

Retraction experiments support Fragmentation. Consider an experiment in which participants were presented with a series of (fake) reports regarding a warehouse fire (Johnson and Seifert 1994). Participants received thirteen messages about the fire, one message at a time. In the fifth message, participants were told that 'cans of oil paint and gas cylinders' were stored in the room where the warehouse fire started. The information was then immediately retracted—participants were told in the next message that there were no oil cans or gas cylinders in the room where the fire started, so they couldn't have caused the fire.[15] After a 10-minute distractor task, participants answered a questionnaire designed to ascertain whether they thought that the oil cans and gas cylinders had started the fire. Participants were asked questions such as: 'Why did the fire spread so quickly?', 'What was the possible cause of the toxic fumes?', and 'Why do you think the fire was particularly intense?'

Perseveration was rampant, with participants seemingly ignoring the retraction and referring to the debunked oil and gas explanations in their responses. Ninety percent of those who saw the correction still made references to the volatile materials in their responses. Some sample responses will give the reader a feel for the results:

Q: Why did the fire spread so quickly? A: 'Oil fires are hard to put out.'

Q: What was the possible cause of the toxic fumes? A: 'Burning paint.'

Q: Why do you think the fire was particularly intense? A: 'The pressurized cylinders.' (Seifert 2002)

One might naturally wonder whether the participants simply missed the retraction, but they did not. For instance, participants were asked, 'Were you aware of any corrections in the reports that you read?' In the initial set of experiments, 100 percent of the delayed correction groups and over 90 percent of the immediate correction groups correctly recalled the retraction (Johnson and Seifert 1994: 1423).

Interestingly, participants appeared to use *both* the initial information and the retraction in inferences. Those in the correction condition averaged 4.9 inferences (out of ten), consistent with belief in the misinformation (Johnson and Seifert 1994: 1424). Other inferences clearly relied on a belief in the correction message

[15] In another condition, participants were given this correction at the very end of the message sequence. Since there were no differences in performance between these two groups, we only discuss the immediate retraction group, as that puts our current point in its starkest contrast.

(participants made references to the correction message or its gist) (Johnson and Seifert 1994: 1424; 2002: 270–271). Assuming that being relied on in (non-hypothetical) inference is sufficient for a state to count as belief, participants believed the misinformation and the correction. In other words, they had contradictory beliefs, and which ones were accessed depended on the type of question being asked. Participants clearly acknowledged the correction and made inferences consistent with it, except when the questions were causal.

Misinformation persistence is, sadly, not exotic. Consider Anderson and Kellam (1992), in which participants were asked to think about the connection between firefighting and risk aversion. This hypothetical reasoning task led to a belief about the connection between the two variables. Participants were then provided (graphical) data that directly contradicted their antecedently formed beliefs. Participants' original beliefs persevered, even though they understood and accepted the data being presented. When asked questions about concrete cases, people used the beliefs they had formed via hypothetical reasoning; when asked about abstract general cases, participants used the beliefs they had formed via the data. This phenomenon is quite general: experimental philosophy is riddled with examples of people answering 'abstract' questions one way and giving contradictory answers to the same questions asked in 'concrete' ways (see Sinnott-Armstrong 2007; Mandelbaum and Ripley 2012).

The Web renders misinformation persistence mysterious, while Fragmentation can explain it. When participants initially hear misinformation, they encode it into one fragment and then encode the correction into a different fragment. Different questions asked at different times can tap into different fragments. The details can be fleshed out in many ways. In the warehouse fire case, participants could have initially encoded the misinformation into a fragment focused specifically on causes of the fire. However, since the correction was not presented as a causal alternative (or merely because the change in temporal context), it could have been encoded in a separate fragment. When asked about causes of the fire, participants would have then accessed information in the first fragment, and when asked about the correction they would have accessed information in the second.

3.2.3. Choice Blindness

People are often blind to their own decisions, even ones they have just made. When asked to recall a recent choice between two options—whether it be between two colors (Lind et al. 2014) or two faces (Johansson et al. 2005)—people can be easily fooled into asserting a belief contrary to fact. For example, consider an experiment in which people are asked to choose which of two faces is more attractive. After the participants have made their choices, the faces disappear for a second and then reappear. The participants are told that the face that reappears was the one they just chose, but sometimes this is inaccurate and the unchosen face appears instead. Participants are then asked to justify their 'choice,' even

though they are presented with the unchosen face. Not only do most people not detect that any switch has been made, but they then offer post hoc rationalizations of their 'choice.'

It's not just attractiveness ratings that involve this kind of blindness—people are often blind to their own political opinions as well. For example, people who agree with the statement 'Even if an action might harm the innocent, it can still be morally permissible to perform it' sometimes fail to realize when 'permissible' is changed to 'forbidden' and even offer post hoc rationalizations of why they think their now extremely different, often contradictory, 'choice' is correct (Hall et al. 2012). Moreover, these effects aren't transient—they affect downstream decision-making in both the present and the future.

Choice blindness brings up a worry: If people's choices are so malleable even regarding supposedly strongly held political opinions, then how can we make sense of there being any persisting stored attitude? Yet, if there are no persisting attitudes, then how can we make sense of attitude polling and the consistency with which people respond to surveys in more 'normal' circumstances? In Hall et al. (2012), participants were asked about pressing national issues that were part of the debate in a very polarized country (Sweden, which is as polarized as America, though shifted left). As Hall et al. say, 'To claim that half the Swedish population holds no articulated attitudes about the most visible moral issues in the current societal debate is a most uninviting conclusion to draw.' So how do we make sense of this situation?

Fragmentation gives us a tool with which to ease the tension. When asked which attitude the subject holds, the answer is: both. Even though the beliefs are inconsistent, both beliefs are under the control of different stimuli. Fragmentation can explain how we store two inconsistent beliefs while still performing rational actions and staving off behavioral paralysis.

3.2.4. Illusory Truth and Fluency

People use fluency as a guide to judge what is true. Sentences that are more fluent (e.g. by being easier to read) are thought to be truer than sentences with the same content but in a more difficult-to-read font (Unkelbach 2007). Similarly, the more frequently one has encountered a sentence, the more fluent that sentence becomes, thus increasing one's rating of the truth of the sentence (a phenomenon sometimes termed 'the illusory truth effect' (Hasher et al. 1977)). But what happens when one encounters information one antecedently knows to be false? The idea that prior knowledge constrains the 'illusory truth' effect is both common sense and a prediction of the Web. Repeating the statement 'The moon is made of green cheese' shouldn't make you believe it if you know it's false.

However, Fazio et al. (2015) have demonstrated that knowledge does not in fact constrain the illusory truth effect. Even statements that participants judge to be false (e.g. 'The Atlantic Ocean is the largest ocean on earth') are susceptible to the

illusory truth effect, with their perceived truth increasing with repeated exposure. As Fazio et al. (2015: 996) put it, '[r]eading a statement like "A sari is the name of the short pleated skirt worn by Scots" increased participants' later belief that it was true, even if they could correctly answer the question "What is the name of the short pleated skirt worn by Scots?"' And ratings didn't just increase—repeated exposures cause participants to judge statements as *true*. Around 50 percent of the known falsehoods were judged to be true after a single repeated exposure (on a binary true/false test).

Illusory truth effects thus provide a separate source of evidence for inconsistent beliefs, which Fragmentation can accommodate and the Web cannot.

3.3. The Crowd Within

A crowd frequently contains more accurate information than any of the individuals that compose it. The average of the guesses of a group are often more accurate than any individual's guesses (Galton 1907; Surowiecki 2004; cf. Navajas et al. 2018). This effect is widely known as 'the wisdom of the crowd.' How a crowd achieves its accuracy is relatively clear: the average of a group's guesses becomes more accurate as the sources of error become more independent. But how individuals generate their guesses in the first place is less clear. Historically, researchers have assumed that people use all of their stored information to guide their judgments, with subsequent judgments just adding noise (Mussweiler et al. 2000). This standard view fits nicely with the Web and also makes a corresponding prediction. Since the estimate of a single person represents the best information available to him or her, additional estimates without interim learning should merely introduce additional chances for error and thus lead to less accurate guesses.

The Web's prediction isn't borne out, however. Surprisingly, one can find the wisdom of the crowd effect within single individuals (Vul and Pashler 2008). For example, participants were asked questions to which no one could be expected to know the exact answer, such as 'What percentage of the world's airports are in the United States?' Participants guessed and were then either immediately asked the same question again or asked the same question after a three-week delay. Neither group was provided with advance notice that they would be answering the questions a second time.[16]

The results are instructive. In both conditions, the error of the average of the two guesses was significantly smaller than the error of either guess alone (and the longer the delay, the stronger the effect). Since second guesses tended to be worse

[16] Participants were asked not to reproduce the same answer, and they were explicitly told that they were not specifying a range.

than the first, we can conclude that participants did not produce a second guess by accessing more accurate information.[17] Second guesses also weren't the result of modifying the total information available to a subject because the effect depends on the error of the guesses' being independent. If the second guess were merely based on information modification, the error would likely not be independent enough to generate the effect. It is unclear how the Web could accommodate this data. If the Web is right, we should have global access to our stored information. If we had global access to our stored information, why would multiple guesses from single subjects systematically exhibit errors with the degree of independence necessary for the effect?

Fragmentation, on the other hand, has the resources to explain the data. Different spatiotemporal contexts causes the participants, when queried, to access different fragments, which contain different information (see section 3.1 for more on this point). Guesses produced by drawing on information in one fragment are likely to have different sources of error than those produced by drawing on information in another fragment. The reason the delay group does better than the immediate group is that the delay reduces the anchoring effect produced by a first guess and allows an unrelated fragment to be activated, which typically increases the independence of the sources of error in the guesses.

So far, we've seen that the Web struggles to explain results that suggest humans store redundant representations, inconsistent beliefs, and beliefs with independent sources of error. Troubles with the Web don't end there. It also can't seem to accommodate a basic result from social psychology: people are extremely resistant to revising positive beliefs about themselves.

3.4. Structuring Beliefs: The Self as the Center of Doxastic Gravity

Consider your reaction to being told that ace mathematicians have invented non-standard algebras where 2 + 2 doesn't necessarily equal 4. Now compare that reaction to the one you would have if you were told that top ethicists have been closely observing your behavior and have concluded that you are a very bad person. Or that excellent psychometricians have determined that you possess below-average intelligence. Whereas your response to the mathematical counter-attitudinal information might be surprise and curiosity, your response to receiving the information about yourself is likely to be far more intense. You can get people to accept the former merely on the advice of experts (just as one can

[17] In our experience, people get confused on this point because, we suspect, they are mistaking the population level effect for one that holds for each specific guess. Not every pair of guesses exhibits this pattern—the pattern holds for the population of guesses (so sometimes first guesses are worse than second ones, but on average they are better). Regardless, the mere fact that second guesses were often mistaken at all is enough to show that subjects weren't looking up the answers.

by telling people that expert logicians have found that modus ponens isn't valid), but it's unlikely that there is any amount of evidence—never mind mere testimonial evidence—that will make people believe they aren't good or smart. This suggests that the belief that you are a good person and the belief that you are a smart person are extremely strong beliefs, which accords with a similar moral stemming from dissonance theory.[18]

Take the classic effort justification paradigm, which shows that those who put lots of effort into a task paradoxically increase the degree to which they like that task. For instance, participants whose initiation into a group requires higher-voltage shocks will like that group more than participants whose initiation requires lower-voltage shocks (Gerard and Mathewson 1966),[19] and participants who have to tell an embarrassing story about themselves to join a group end up liking the group more than those who don't have any such painful barrier to joining (Aronson and Mills 1959).

Standard explanations of the effect invoke cognitive dissonance, an unpleasant sensation induced by situations involving inconsistent beliefs and behaviors. The participants experience dissonance during the experiment and unconsciously change their attitude in order to alleviate the dissonance.

What induces dissonance? Somewhat controversially (see, e.g. Cooper 2007), we hold that dissonance is generated when people become aware of having inconsistent attitudes. In early dissonance theorizing, all dissonance was taken to be caused by logically inconsistent attitudes. Unfortunately, this simple proposal has a serious problem as mere logical inconsistency doesn't explain paradigmatic cases of dissonance, e.g. someone's believing that smoking is bad for them and still smoking. At least at first glance, there isn't any logical inconsistency between believing one is a smoker and believing smoking is bad. A natural thought, then, is that the simple theory of the cause of dissonance must be wrong. Dissonance must be generated by more than mere logical inconsistency.

However, if dissonance isn't triggered solely by logically inconsistent attitudes, then dissonance theory is in trouble. Positing dissonance arousal quickly becomes unprincipled and ad hoc (e.g. any attitude change whatsoever could in principle be chalked up to dissonance reduction).[20] It would be a theoretical coup if we could

[18] To clarify, these beliefs need not be conscious and, because our model allows for inconsistent beliefs, needn't imply that people don't also hold the opposite beliefs. Also, we suppose that holding these core beliefs is a central tendency for people, but we needn't deny individual differences. For one thing, individual differences are needed anyway to track differences in sensitivity to dissonance (Heitland and Bohner 2010); for another, some people truly just think they are bad people (and this is a group that is often categorized as non-neurotypical and tends to suffer from depression).

[19] Note that this effect only holds if the shock is a means of joining the group. If it's just incidental to the group, or if one gets shocked and just evaluates the group, then the effect doesn't hold. This helps to show that the effect is propositional rather than associative—the activation spreading of the negative association can only get a foothold on attitudes when the dissonance reasoning isn't at play.

[20] Hence the old joke: if you want to know what situations cause dissonance, just ask Leon (Aronson 1992).

find some general rule to explain when dissonance will arise. We think there is an easy way to save the idea that the root cause of dissonance is logical inconsistency while still explaining why dissonance is generated in cases that don't seem to involve any obvious logical inconsistencies. The explanation for how to do so presupposes fragmentation, so to the extent that the explanation is fruitful, we find more support for the view.

If one assumes that people have *core beliefs* that they are good people, smart people, and reliable, consistent, strong people (Thibodeau and Aronson 1992), one can find logical inconsistencies in all dissonance paradigms. Let's walk through a particular case of effort justification. Suppose a person joins the Marines and undergoes a time-consuming and painful initiation ritual. Surprisingly, the harsh initiation doesn't make her dislike the Marines. In fact, she likes it even more than she would if the initiation were less time-consuming and painful. Dissonance theory says that the increased liking is caused by an attempt to reduce dissonance. But what inconsistency could be causing the dissonance in the first place?

This is where the notion of core beliefs arises. One reasons in the following way:

P1) I put a lot of effort into joining the Marines.
P2) Only an idiot would put a lot of effort into joining the Marines without liking the Marines.
P3) I am not an idiot.
C) I must like the Marines. (And the appraisal of the group is thus increased.)

Evidence for this type of (unconscious—see Lieberman et al. 2001) reasoning comes from a variety of sources. One worth mentioning: if you negate one of these premises, the effect vanishes. Participants who have lower self-esteem (and are thus predisposed to think they are stupid) don't show the normal effort justification effect. This happens regardless of whether low self-esteem is manipulated (Glass 1964; Aronson and Mettee 1968) or trait based (as in depression; Rhodewalt and Agustsdottir 1986). That makes sense when understanding dissonance-based attitude change as a conclusion of unconscious reasoning—remove a premise and the modus tollens falls apart.

A few morals: first, the logical inconsistency is what generates dissonance and initiates the appraisal. It is the desire to avoid believing that one is and is not an idiot that initiates the process of belief change. Second, this kind of reasoning is totally general—it has nothing to do with the Marines and everything to do with keeping up one's sense of one's own competence. The same style of reasoning can be found in any of the traditional dissonance paradigms and with any type of decision.

This brings us to a few questions. Why do core beliefs about the self appear to be so strong? And why do core beliefs appear to be active in post-decisional

processes regardless of the dissonance paradigm or the content of the decision? Why do beliefs about one's self-image keep cropping up regardless of whether one is deciding what appliance to buy, what group to join, where to eat, or whom to marry?

Our use of 'core beliefs' is just a variant of the idea of the self as a schematic trait that is so important that it colors all information processing (Markus 1977) because one's core beliefs are chronically accessible (Linville 1985). That is, any evidence about anything can be taken, first and foremost, to reflect how one views oneself. In sum, one's self-image dictates how one restructures one's beliefs (Mandelbaum 2019).[21]

According to the picture on offer, one's self-concept is extremely representationally redundant. Fragmentation interprets the strength of core beliefs as, in part, a function of how many copies of a given belief one has. The more redundant a belief is, the more contents it will be related to, which increases its inferential promiscuity. Core beliefs are hypothesized to be extremely redundant and thus should be recruited in reasoning about all sorts of disparate categories.[22] At the limit, these beliefs would be 'most central' by being maximally redundantly represented and appearing in all belief fragments.[23]

Insofar as dissonance theory needs core beliefs—and we think they're crucial to understanding the mechanism of dissonance theory—the Web's conservatism requirement runs afoul of them. Conservativism says the beliefs that are least open to revision are those that would require the highest number of changes. But surely changing beliefs about the rules of logic or arithmetic would require a higher number of changes than changing your belief that you are a good person. We think this is true no matter how one counts belief changes, since if you revise your beliefs about modus ponens, you would have to make an infinite number of

[21] It may seem implausible that self-beliefs are activated when you are thinking about matters as lofty and esoteric as quantum mechanics or fine art (or, for that matter, as quotidian as shooting a free throw or smoking a cigarette [Mandelbaum 2020]), but what it is for a representation to be chronically accessible is for the representation to be activated across contexts, even wholly unrelated ones. There's no logical reason why core beliefs should arise in thinking about black holes or mellifluous birds, which is what makes the psychological fact that much more interesting. The power of human narcissism knows no bounds. The idea that beliefs about the self are ubiquitous in belief updating—whether this is because they're explicitly represented in every fragment or whether they're built into the architecture (Quilty-Dunn and Mandelbaum 2018b) and implicitly represented—isn't just supported by data on chronic accessibility, it's also presupposed by effort justification. The argument goes something like this: dissonance is the best explanation of effort justification; effort justification occurs across any content domain; effort justification relies on the existence of core beliefs; in a fragmented architecture, accessibility is just a matter of redundant representation (or rules built into the architecture), so core beliefs must be embedded in each fragment (or built into the architecture).

[22] It also entails that any information can be deemed self-relevant, which sadly concurs with informal observation.

[23] This isn't to say that one's conception of the self is consistent. Those high in "self-complexity" will have very different conceptions of themselves depending on the content domain, whereas those lower in self-complexity will be more consistent throughout content domains (see, e.g., McConnell et al. 2009). Fragmentation can easily model this distinction, for those high in self-complexity are just people with inconsistent beliefs, where the beliefs are about the self.

revisions to maintain consistency. Fragmentation does not interpret belief centrality as a function of logical strength and thus can accommodate beliefs about the self as the central doxastic point from which all other belief change pivots.

* * *

The primary upshot of the discussion so far is that there are serious problems afoot for the Web. People do not update globally. Their total set of beliefs is replete with redundancy and inconsistency. Their contingent, often false beliefs about themselves are strongest, while beliefs about logical and mathematical truths don't appear to be particularly strong.

Fragmentation thus better coheres with human psychology than the Web. Fragmentation is still just a framework, however. Below, we offer some speculative principle to flesh out the picture and set the groundwork for the beginning of a theory of the architecture of belief storage.

4. The Functional Profile of Fragments

To provide a theory of fragments is, in part, to provide a theory of their functional role. This requires answering questions such as 'How are fragments created and destroyed?' and 'What determines which fragments are activated?' We attempt to answer these questions by providing a saliency structure for fragments along the following three basic lines: the environment the information was acquired in, the amount of information the fragment contains, and how recently the fragment was activated.

Work on the structure of belief storage is in its infancy. What we need at this point are hypotheses that aren't obviously ruled out by intuition and empirical data, and perhaps even supported by them. What follows is one such hypothesis about how fragment activation, creation, and interaction could work.

4.1. Fragment Activation

A lesson we can draw from the considerations presented above is that fragments appear to be organized around the environment in which the information was acquired. Recall that spatiotemporal context is the primary determinant of which stored information is accessed by subjects in the associative renewal, wisdom of the crowd, and implicit bias case studies.

To account for the role of spatiotemporal context in information access, we propose that fragments, like mental files (Fodor 2008), are searched via their headings. Conceptualizations of environmental scenarios act as headings for

fragments and information acquired in a particular setting is stored within the corresponding fragment. Suppose you enter a new cafe, Bar Bert, for the first time. A new fragment will be opened, and the fragment name will correspond to your concept of the cafe. Any new information learned while in the cafe will be stored in the corresponding fragment. Say that while in the cafe you learn a new word: 'pusillanimous.' Further, assume that you come to believe that cowards are pusillanimous and that 'pusillanimous' is a prime example of a needlessly pretentious synonym. Both beliefs will be stored under the BAR BERT[24] heading, so that you will be quicker to recall that cowards are pusillanimous if you are first reminded of Bar Bert. Similarly, if asked for an example of a needlessly pretentious synonym, recovering PUSILLANIMOUS will activate all of the other information in the BAR BERT fragment.[25]

This brings us to the first principle of fragment activation, the *Headings Principle*: in the first instance, exogenous stimuli trigger searches through fragment headings, not fragment information, for information that matches the stimuli. Imagine that after leaving Bar Bert, you go to a new coffee shop, Cafe Cat, and meet a friend who tells you new information about Bar Bert. You would then open a CAFE CAT fragment, and in it would be the new information about Bar Bert. If at a later time someone mentioned Bar Bert to you, you would, ceteris paribus, open up the BAR BERT fragment and not the CAFE CAT one, even though the latter may have more information about Bar Bert than the former.

The Headings Principle helps to explain more quotidian facts about recall as well. For example, people are more apt to remember a piece of information if they are in the same situation in which they learned that information. More generally, information learned in a given context is more successfully retrieved when retrieval occurs in that same context (Thomson and Tulving 1970). This effect is sometimes called 'context-dependent memory' (Smith and Vela 2001) and at other times the 'encoding specificity phenomenon.' This makes sense on a view where information is stored underneath headings pertaining to the environment in which it is learned.

More evidence for the Headings Principle comes from the 'walking through doorways causes forgetting' effect. Merely leaving a room appears to trigger forgetting: if you are holding an object while walking through one room, you

[24] For simplicity's sake we assume that concepts serve as fragment headings, though as far as we can see nothing hangs on this.

[25] Caveat: we have been mostly discussing beliefs throughout the chapter, not concepts. One might wonder whether there is a central non-fragmented database of concepts. Though we find this question fascinating, we are agnostic about this. Phenomena like semantic priming appear to suggest an affirmative answer, but priming can be understood in different ways. For example, there is evidence that the language parser has access to a web-like lexicon (see Mandelbaum 2015b). Whether its elements count as concepts—and whether they are available for use by any other process or can explain priming—are interesting questions. If they cannot, perhaps priming suggests a separate, central, globally accessible store of concepts.

are more likely to forget what you are holding if you enter a different room (Radvansky and Copeland 2006). That is, people remember the object more when they are probed in the same room in which they picked it up, even when distance traveled, length of time, etcetera are controlled for. The idea is that when a person walks through a doorway, an event boundary is crossed. On the assumption that there can only be one event model active at a time (Radvansky et al. 2011), when one updates one event, the last one becomes inactive. The events in Radvansky et al. are, in our lingo, just triggers for fragments: ceteris paribus, one creates a new fragment for each new room one enters (for more on how the environment dictates fragment genesis, see the next section).

But what about the commonplace cases where new information is acquired in a familiar setting? In the first instance, the familiar setting will activate the fragment for that location, and thus activation will proceed as discussed. Imagine that you become a regular at Bar Bert and learn that the warbler spring migration passes through New York in May. That information will be stored in the BAR BERT fragment.

However, if the information you learn about warblers is important, it could be useful for that information to be separate from one's knowledge about Bar Bert. Under conditions of load, such information stays put, but for subjectively important information (say, one would like to see more warblers), the information can be copied into a new fragment, perhaps one headed by WARBLER. The idea is that although in the first instance fragment names are places, any concept could in principle become a fragment heading because of its subjective value.

What if you see warblers in a familiar environment, and thus aren't likely to open a new fragment, but have no preexisting WARBLER-headed fragment to activate? You could default to another salient, detected property to dictate where new information is stored. Call this principle the *Secondary Property Principle.* If there is no WARBLER-headed fragment that has recently been activated, then (assuming you see a tree too) you could search for a TREE-headed fragment instead and put the information there.

But how would fragment activation work if you are not in a novel environment, there is no fragment headed by WARBLER, and there is no detection of a salient secondary property? We hypothesize that in this case a fragment that has been activated very recently will be the most likely to be reactivated. Call this the *Recency Principle.* Fragments that have recently been open will be more salient than those that haven't and so will be more apt to be activated. Newly acquired information will thus tend to be added to recently activated fragments.

Finally, if there is no salient secondary property, appropriate heading, or very recently opened fragment, we assume that the search defaults to the largest fragment. 'Largest' here means fragment that contains the highest number of preexisting links between distinct beliefs (and thus in some sense, the fragment will contain the most information). Call this the *Size Principle.*

How do these principles interact? Let's walk through a hypothetical case. Suppose you are looking at birds and wonder when warblers pass through New York. We expect that the first search will proceed through the currently active fragment, which, ceteris paribus, will be antecedently determined by your environment. If that search is unsuccessful, then you will search for a fragment headed by WARBLER. If none is found, the search will proceed through a fragment headed by one of the other detected properties, such as TREE (or even NEW YORK, since that is a constituent of what you are wondering), reactivating the most recently activated fragment headed by the secondary property. Finally, if the search has still not succeeded, you will default to the largest fragment.

4.2. Fragment Genesis

Now that we've explained how new information might be added to preexisting fragments, and how fragments might be searched, we can turn to how fragments could be created in the first place and how they might interact. Below we sketch a model of fragment genesis and propose a principle governing fragment interaction.

We hypothesize that fragment creation is governed by the *Environmental Principle*: novel fragments are opened up in novel environments. According to this principle, when one visits Spain for the first time, one opens up a new fragment with SPAIN as the heading. Of course, one doesn't just visit Spain; one goes to the Madrid Airport or the *Sagrada Familia*. For each of these places, we assume that a new fragment will be opened. One's Spain fragment is unlikely to encompass all the information one encodes about Spain, since one is unlikely to activate SPAIN in every environment in Spain. Indeed, we suspect that the fragment headed by SPAIN will contain less information than the fragments pertaining to the particular cities in Spain. The situation is analogous to what happens with basic-level concepts. Since BIRD is a basic-level concept, it is more informationally rich than ANIMAL, even though all birds are animals (Rosch 1978).[26] Similarly, even though all information about Madrid is also information about Spain, the more specific place takes precedent. Particular places one goes to in a city—the Alhambra, or Bar Bert—will have their own autonomous information store.

The Environmental Principle is supported by data from the memory reconsolidation literature. Hupbach et al. (2008) had participants memorize a list of twenty everyday objects (e.g. a tennis ball, a stapler, a toy car, etc.), which were placed in a certain distinguishing learning context (e.g. a blue basket). Later,

[26] For discussion on why this holds for perceptual reasons, see Mandelbaum (2018).

participants were asked to remember a second list of semantically unrelated everyday objects. These objects were merely laid out on a table, thus distinguishing this second context from the first. One group of participants was reminded of the first list before learning the second, while another wasn't. The reminder group was asked what they remembered about the first learning task and were then brought back to the same room as the first task to learn the second list; the no-reminder group learned the second list in a different room, with a different experimenter. The initial test itself was a free recall, in which participants were asked to write down all the items from both lists sorted by day, but subsequent versions of the experiment were successfully conducted with recognition tasks.

The most important finding for our purposes was an asymmetry in the participants' mistakes. Participants in the reminder condition mistakenly recalled a large number of items from the second list when they were asked to recall items from the first list. However, they did not mistakenly recall items from the first list when asked to recall items from the second list. Participants in the no-reminder condition exhibited significantly fewer intrusions, and the intrusions that did occur were symmetric.

The Environmental Principle can explain this effect. List one is always learned in a novel context, so it opens up its own memory fragment. Presenting a reminder of list one activates the memory of list one. When items from list two are presented after the reminder, these items are used to update the newly malleable list-one fragment. Later, when participants are asked to recall items from list one, they recall both the correct items and items from list two which were mistakenly entered into the list-one fragment. Participants in the no-reminder condition have no such activation of the first fragment when learning the second list. Because changes in spatiotemporal context open new fragments, shifting rooms creates a new fragment in which the second list can be stored—hence the low level of intrusions going in either direction.[27]

4.3. Fragment Synchronization and Merge

Finally, we come to the question of how fragments synchronize and merge. This question is intimately bound up with how a fragmented mind deals with activated inconsistencies. How it does will depend on whether the inconsistency

[27] For those still keeping score, the Web lacks the resources to accommodate such patterns because of its commitment to global updating. A single non-redundant, continually globally updated information store should display no intrusion effects at all (beyond noise). However, the primary problem for the Web isn't the errors per se but the asymmetry in the errors. As far as the Web is concerned, one should predict that it is just as likely for memories of list two to bleed into memories of list one as it is for memories of list one to bleed into memories of list two.

is intrafragmental or interfragmental. Although there is intrafragment consistency, interfragment consistency isn't automatically maintained. Suppose a single fragment contains both the belief that P and the belief that $\neg P$. This would be intrafragment inconsistency and would be automatically resolved. After inconsistency detection, processes unfold along the lines predicted by dissonance theory. Inconsistencies cause dissonance, which is a specific type of phenomenologically salient discomfort. Because it is uncomfortable, people tend to alleviate dissonance as soon as they can. Dissonance assuagement can take many different forms. One may try to focus on the discomfort as opposed to its underlying cause, for example by shifting one's focus to what (orthogonal) property one really values (Tesser and Cornell 1991); one might try to focus on the phenomenology to take away the acuteness of the discomfort (Ochsner and Gross 2005); or one might simply forget all that and just get drunk (Steele et al. 1981). But these are all temporary salves.[28] The underlying inconsistency will still lurk, and dissonance will reappear.

A more direct route is to deal with the inconsistency head-on, changing one of the offending attitudes to alleviate the dissonance. One could perhaps erase the belief that one has less evidence for (or more likely, the one that one identifies with less).[29] Otherwise, one could sequester one of the two inconsistent beliefs by sending it to a different fragment. Sequestering a belief severs its inferential connections, making it less likely to become reactivated.

Interfragment inconsistency is handled differently. In this case, one might believe P and believe $\neg P$ but experience no dissonance because P is in an activated fragment while $\neg P$ is not. Since no dissonance would be created, there would be no motivation to restructure one's beliefs. However, imagine a case where there is a reminder that serves to activate a previously quiescent fragment that contains $\neg P$. Here, one would have two activated fragments, one containing P and one containing $\neg P$. In this case, we expect the fragments to merge: the fragments will combine their information while deleting (or sequestering) the weaker belief.

The general principle *Merge* is as follows: if two fragments that contain inconsistent information are coactivated, then they will be rendered consistent. Information that leads to inconsistency is either deleted, deactivated, or

[28] There must be some duration after which active fragments become inactive. Were one to pull one's attention away from the inconsistency for long enough, one might be able to have time to deactivate the fragment. This would push the dissonance off considerably; the salve would still be temporary, but the difference would be in days or weeks, not in seconds or minutes. For what it's worth, there is evidence that priming mindsets of trust and distrust can switch from trial to trial (with 800 ms exposures!), thus allowing for the possibility that fragmentation activation and quiescence might occur within seconds or less (Schul et al. 2004).

[29] We are skeptical of the idea of ever erasing beliefs, at least for cognitive reasons such as no longer having evidence for the belief or having a second-order belief that one shouldn't hold the belief (of course, we still believe that you can lose a belief by intervening on variables one level down—e.g. getting hit in the head with a rock).

sequestered. The two previous fragments are then merged to form a new, consistent fragment.[30]

Although *merge* is speculative, there is some suggestive evidence in its favor. After attitude change, people have trouble recovering their earlier attitude, even if, as often occurs, it was recently offered. Moreover, when forced to guess at what their previous attitude was, people tend to think their old attitude was identical to their current attitude (Bem and McConnell 1970; Ross 1989; Lieberman et al. 2001; see Hall et al. 2012 for an interestingly related effect).

4.4. Inconsistent Beliefs and Practical Action

While the principles described above help to explain the previously discussed data, they do nothing to explain how one can engage in practical reasoning if one believes inconsistent propositions. It is clear, however, that people do engage in practical reasoning and that they have inconsistent beliefs. So how do they do it?

One main way is by decreasing the likelihood that inconsistent beliefs will be coactivated. Fragmentation allows for the sequestering of inconsistency, but how does the mind actually reduce the likelihood of coactivating inconsistent beliefs? Perhaps the mind accomplishes this by operating in accordance with the 'let sleeping dogs lie' principle (McDermott 1987). Roughly, the principle is one of cognitive economy: one conserves cognitive energy unless spurred on by an external event or command. Applied to Fragmentation, the principle dictates that a fragment remains quiescent unless (a) a search is triggered for its specific heading, and (b) once that heading is located, searches cease. As long as inconsistent beliefs are housed in separate fragments, a sleeping-dogs principle dramatically decreases the likelihood of coactivating the inconsistent beliefs. A mind that follows such a principle thereby allows for practical reasoning in the face of inconsistent information storage.

5. Conclusion: The End of the Beginning

The principles outlined above mark the very beginning of the construction of a theory of the cognitive architecture of belief storage. Although the theory might seem radical, it has much in common with several models of memory.[31] Memory

[30] *Merge* isn't the only option available. One might instead want to have the possibility of synchronization without merge, whereby two fragments synchronize but keep their contents as is and continue along their own trajectories, allowing for future divergence.

[31] See the work being done in the laboratory of Ken Norman (e.g. Baldassano et al. [2017]), for some parade examples of what look like fragmented models we propose in the neuroscientific study of memory. Other computational models of memory also share deep similarities with our flavor

researchers have long held that there are many different stores of memory, e.g. episodic vs. semantic memory. Additionally, redundant representations could be interpreted as a type of memory trace, as on the multiple-trace hypothesis (Hintzman and Block 1971). The Environmental Principle just falls out of work from context-dependent memory (Shin et al. 2020). Even the idea that memory is composed of content-neutral, isolated data structures accessible only via their headings can be found in headed records and some schema models of memory (Jones 1984; Morton et al. 1985).

Fragmentation, as offered here, suggests specific process-level generalizations about how information is stored and accessed, and thus our project differs substantively from those of Lewis (1982), Stalnaker (1984), Egan (2008), Rayo (2013), Yalcin (2018), and Elga and Rayo (Chapter 1 in this volume).[32] These theorists are agnostic about how beliefs are stored and processed in human minds (not coincidentally, none of them are representationalists about belief). They are content with a fragmented picture of belief's being a modeling convenience instead of a hypothesis about cognitive architecture.[33]

Whatever one's views on the nature of belief or the psychological reality of Fragmentation, one thing is clear: there remains much work to be done on the content, architecture, and implementation of belief. We have argued that different models of the architecture of belief storage generate different predictions about cognition and that Fragmentation's predictions are more consistent with the best available evidence than the Web's. We have not established that the Web is false or that Fragmentation is true. In fact, we suspect that the truth lies somewhere between both extremes. However, to begin to discover the structure of belief storage, we need proposals specific enough to generate predictions and be productively developed in the light of new countervailing evidence. We think Fragmentation is such a proposal, but whatever the structure of belief storage, we hope others will join us in mapping this vast terra incognita.[34]

of fragmentation. For example, the model constructed by Ecker et al. (2011), for instance—a type of stochastic sampling model—presupposes representational redundancy, non-global updating, and the storage of inconsistent information, which makes it an instance of Fragmentation.

[32] Cherniak (1981, 1983), on the other hand, did intend to offer a fragmented model of cognitive architecture (which he called 'compartmentalization'), but his model was sparse on details and process-level generalizations. His primary aim seemed to be to explain how we can understand the rationality of a human agent's actions. Nevertheless, our work here can be construed as an attempt to extend the descriptive side of that type of project.

[33] Even though we doubt that any of these theorists would endorse the view we are calling Fragmentation, there is surely common ground between us all. For instance, Fragmentation puts certain constraints on the logic of belief that anyone who rejected the Web would likely accept. Given almost any version of a fragmented view, belief will not be closed under conjunction introduction or known entailment.

[34] This paper has been circulating for a very long time, and as such there are a wide variety of people, institutions, and drinks to thank. Help and support came from so many venues that our list is surely woefully incomplete, but at the very least we must thank Ned Block, Andy Egan, Adam Elga, Tatiana Emmanouil, Dan Harris, Zoe Jenkin, Bence Nanay, Ian Phillips, Nic Porot, Jake Quilty-Dunn, Agustín Rayo, David Rosenthal, Susanna Siegel, Pepa Toribio, students in EM's and Ned's perception seminar,

References

Anderson, C., and Kellam, K. (1992), 'Belief Perseverance, Biased Assimilation, and Covariation Detection: The Effects of Hypothetical Social Theories and New Data', *Personality and Social Psychology Bulletin* 18/5: 555–565.

Aronson, E. (1959), 'The Effect of Severity of Initiation on Liking for a Group', *The Journal of Abnormal and Social Psychology* 59/2: 177–181.

Aronson, E. (1968), 'Dishonest Behavior as a Function of Differential Levels of Induced Self-Esteem', *Journal of Personality and Social Psychology* 9/2, pt.1: 121–127.

Aronson, E. (1992), 'The Return of the Repressed: Dissonance Theory Makes a Comeback', *Psychological Inquiry* 3/4: 303–311.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017), 'Discovering Event Structure in Continuous Narrative Perception and Memory', *Neuron* 95(3): 709–721.

Bem, D., and McConnell, H. K. (1970), 'Testing the Self-Perception Explanation of Dissonance Phenomena: On the Salience of Premanipulation Attitudes', *Journal of Personality and Social Psychology* 14/1: 23–31.

Bendaña, J. (2021), 'Implicit Attitudes are (Probably) Beliefs', in C. Borgoni, D. Kindermann, & A. Onofri (eds), *The Fragmented Mind* (Oxford: Oxford University Press).

Bouton, M. (2004), 'Context and Behavioral Processes in Extinction', *Learning & Memory* 11/5: 485–494.

Bustamante, J., Uengoer, M., and Lachnit, H. (2016), 'Reminder Cues Modulate the Renewal Effect in Human Predictive Learning', *Frontiers in Psychology* 7: 7, 80. URL: https://doi.org/10.3389/fpsyg.2016.01968.

Cherniak, C. (1981), 'Minimal Rationality', *Mind* 90/358: 161–183.

Cherniak, C. (1983), 'Rationality and the Structure of Human Memory', *Synthese* 57/2: 163–186.

Churchland, P. (1981), 'Eliminative Materialism and the Propositional Attitudes', *The Journal of Philosophy* 78/2: 67–90.

Cooper, J. (2007), *Cognitive Dissonance: 50 years of a Classic Theory* (Sage).

De Houwer, J. (2009), 'The Propositional Approach to Associative Learning as an Alternative for Association Formation Models', *Learning & Behavior* 37/1 (2009): 1–20.

students in the Ignorance and Stupidity seminar, and audiences at the University of Antwerp, the University of Barcelona, TU Dortmund, the Jean Nicod Institute, Marist College, the University of Michigan, Princeton University, and the European Society for Philosophy and Psychology. Thanks also to the National Endowment of the Humanities for support stemming from award FEL-257901-18, granted to EM, and to the PSC-CUNY for award #62497-00 50, both granted to EM at some point during the elephantine gestation of this paper.

De Houwer, J. (2019), 'Moving Beyond System 1 and System 2: Conditioning, Implicit Evaluation, and Habitual Responding Might Be Mediated by Relational Knowledge', *Experimental Psychology* 66/4: 257–265.

Dennett, D. (1991), 'Real Patterns', *The Journal of Philosophy* 88/1: 27–51.

Devine, P., Forscher, P., Austin, A., and Cox, W. (2012), 'Long-Term Reduction in Implicit Race Bias: A Prejudice Habit-Breaking Intervention', *Journal of Experimental Social Psychology* 48/6: 1267–1278.

Ecker, U., Lewandowsky, S., Swire, B., and Chang, D. (2011), 'Correcting False Information in Memory: Manipulating the Strength of Misinformation Encoding and Its Retraction', *Psychonomic Bulletin & Review* 18/3: 570–578.

Egan, A. (2008), 'Seeing and Believing: Perception, Belief Formation and the Divided Mind', *Philosophical Studies* 140/1: 47–63.

Elga, A., and Rayo, A. (2021), 'Fragmentation and Information Access', in C. Borgoni, D. Kindermann, & A. Onofri (eds), *The Fragmented Mind* (Oxford: Oxford University Press).

Fazio, L., Brashier, N., Payne, B. K., and Marsh, E. (2015), 'Knowledge Does Not Protect against Illusory Truth', *Journal of Experimental Psychology: General* 144/5: 993–1002.

Fazio, L. K., Rand, D. G., and Pennycook, G. (2019), 'Repetition Increases Perceived Truth Equally for Plausible and Implausible Statements', *Psychonomic Bulletin & Review* 26/5: 1705–1710.

Fodor, J. (1983), *The Modularity of Mind* (MIT Press).

Fodor, J. (1998), *Concepts: Where Cognitive Science Went Wrong* (Oxford University Press).

Fodor, J. (2008), *LOT 2: The Language of Thought Revisited* (Oxford University Press).

Galton, F. (1907), 'Vox Populi', *Nature* 75: 450–451.

Garcia-Marques, T., Silva, R., Reber, R., and Unkelbach, C. (2015), 'Hearing a Statement Now and Believing the Opposite Later', *Journal of Experimental Social Psychology* 56: 126–129.

Gerard, H., and Mathewson, G. (1966), 'The Effects of Severity of Initiation on Liking for a Group: A Replication', *Journal of Experimental Social Psychology* 2/3: 278–287.

Gilbert, D., Krull, D., and Malone, P. (1990), 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information', *Journal of Personality and Social Psychology* 59/4: 601–613.

Gilbert, D., Tafarodi, R., and Malone, P. (1993), 'You Can't Not Believe Everything You Read', *Journal of Personality and Social Psychology* 65/2: 221–233.

Glass, D. (1964), 'Changes in Liking as a Means of Reducing Cognitive Discrepancies between Self-Esteem and Aggression', *Journal of Personality* 32/4: 531–549.

Gweon, H., Young, L., and Saxe, R. (2011), 'Theory of Mind for You, and for Me: Behavioral and Neural Similarities and Differences in Thinking about Beliefs of the Self and Other', *Proceedings of the Annual Meeting of the Cognitive Science Society* 33/33: 2492–2497.

Hall, L., Johansson, P., and Strandberg, T. (2012), 'Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey', *PloS ONE* 7/9: e45457.

Hasher, L., Goldstein, D., and Toppino, T. (1977), 'Frequency and the Conference of Referential Validity', *Journal of Verbal Learning and Verbal Behavior* 16/1: 107–112.

Heitland, K., and Bohner, G. (2010), 'Reducing Prejudice via Cognitive Dissonance: Individual Differences in Preference for Consistency Moderate the Effects of Counter-Attitudinal Advocacy', *Social Influence* 5/3: 164–181.

Herrnstein, R. (1969), 'Method and Theory in the Study of Avoidance', *Psychological Review* 76/1: 49–69.

Hintzman, D., and Block, R. (1971), 'Repetition and Memory: Evidence for a Multiple-Trace Hypothesis', *Journal of Experimental Psychology* 88/3: 297–306.

Hupbach, A., Hardt, O., Gomez, R., and Nadel, L. (2008), 'The Dynamics of Memory: Context-Dependent Updating', *Learning & Memory* 15/8: 574–579.

Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005), 'Failure to Detect Mismatches between Intention and Outcome in a Simple Decision Task', *Science* 310/5745: 116–119.

Johnson, H., and Seifert, C. (1994), 'Sources of the Continued Influence Effect: When Misinformation in Memory Affects Later Inferences', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20/6: 1420.

Jones, G. (1984), 'Fragment and Schema Models for Recall', *Memory & Cognition* 12/3: 250–263.

Karlan, B. (2020), 'Rationality, Bias, and Mind: Essays on Epistemology and Cognitive Science', Doctoral Dissertation, Princeton University ProQuest Dissertations and Theses Global.

Lai, C., Skinner, A., Cooley, E., Murrar, S., Brauer, M., Devos, T., . . . and Simon, S. (2016), 'Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time', *Journal of Experimental Psychology: General* 145/8: 1001–1016.

Legare, C., Evans, E. M., Rosengren, K., and Harris, P. (2012), 'The Coexistence of Natural and Supernatural Explanations across Cultures and Development', *Child Development* 83/3: 779–793.

Levy, N., and Mandelbaum, E. (2014), 'The Powers That Bind: Doxastic Voluntarism and Epistemic Obligation', in J. Matheson and R. Vitz (eds.), *The Ethics of Belief* (Oxford University Press), 15–32.

Lewandowsky, S., and Kirsner, K. (2000), 'Knowledge Partitioning: Context-Dependent Use of Expertise', *Memory & Cognition* 28/2: 295–305.

Lewandowsky, S., Roberts, L., and Yang, L. (2006), 'Knowledge Partitioning in Categorization: Boundary Conditions', *Memory & Cognition* 34/8: 1676–1688.

Lewis, D. (1982), 'Logic for Equivocators', *Noûs* 16/3: 431–441.

Lieberman, M., Ochsner, K., Gilbert, D., and Schacter, D. (2001), 'Do Amnesics Exhibit Cognitive Dissonance Reduction? The Role of Explicit Memory and Attention in Attitude Change', *Psychological Science* 12/2: 135–140.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., and Johansson, P. (2014), 'Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say', *Psychological Science* 25/6: 1198–1205.

Linville, P. (1985), 'Self-Complexity and Affective Extremity: Don't Put All of Your Eggs in One Cognitive Basket', *Social Cognition* 3/1: 94–120.

McConnell, A., Rydell, R., and Brown, C. (2009), 'On the Experience of Self-Relevant Feedback: How Self-Concept Organization Influences Affective Responses and Self-Evaluations', *Journal of Experimental Social Psychology* 45/4: 695–707.

McDermott, D. (1987), 'We've Been Framed: Or, Why AI Is Innocent of the Frame Problem', in Z. Pylyshyn (ed.), *The Robot's Dilemma* (Greenwood Publishers), 113–122.

McGregor, I., Newby-Clark, I., and Zanna, M. (1999), '"Remembering" Dissonance: Simultaneous Accessibility of Inconsistent Cognitive Elements Moderates Epistemic Discomfort', in E. Harmon-Jones and J. Mills (eds.), *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology. Science Conference Series* (APA), 325–353.

Mandelbaum, E. (2014), 'Thinking Is Believing', *Inquiry* 57/1: 55–96.

Mandelbaum, E. (2015a), 'Associationist Theories of Thought', *Stanford Encyclopedia of Philosophy* (Summer 2017 edition). URL = https://plato.stanford.edu/archives/sum2017/entries/associationist-thought.

Mandelbaum, E. (2015b), 'The Automatic and the Ballistic: Modularity beyond Perceptual Processes', *Philosophical Psychology* 28/8: 1147–1156.

Mandelbaum, E. (2016), 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias', *Noûs* 50/3: 629–658.

Mandelbaum, E. (2018), 'Seeing and Conceptualizing: Modularity and the Shallow Contents of Vision', *Philosophy and Phenomenological Research* 97/2: 267–283.

Mandelbaum, E. (2020). 'Assimilation and Control: Belief at the Lowest Levels', *Philosophical Studies* 177(2): 441–447.

Mandelbaum, E. (2019), 'Troubles with Bayesianism: An Introduction to the Psychological Immune System', *Mind & Language* 34/2: 141–157.

Mandelbaum, E., and Quilty-Dunn, J. (2015), 'Believing without Reason, or: Why Liberals Shouldn't Watch Fox News', *The Harvard Review of Philosophy* 22: 42–52.

Mandelbaum, E., and Ripley, D. (2012), 'Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The NBAR Hypothesis', *Review of Philosophy and Psychology* 3/3: 351–368.

Markus, H. (1977), 'Self-Schemata and Processing Information about the Self', *Journal of Personality and Social Psychology* 35/2: 63–78.

Mitchell, C., De Houwer, J., and Lovibond, P. (2009), 'The Propositional Nature of Human Associative Learning', *Behavioral and Brain Sciences* 32/2: 183–198.

Morton, J., Hammersley, R., and Bekerian, D. (1985), 'Headed Records: A Model for Memory and Its Failures', *Cognition* 20/1: 1–23.

Mussweiler, T., Strack, F., and Pfeiffer, T. (2000), 'Overcoming the Inevitable Anchoring Effect: Considering the Opposite Compensates for Selective Accessibility', *Personality and Social Psychology Bulletin* 26/9: 1142–1150.

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., and Sigman, M. (2018), 'Aggregated Knowledge from a Small Number of Debates Outperforms the Wisdom of Large Crowds', *Nature Human Behaviour* 2: 126–132.

Newby-Clark, I., McGregor, I., and Zanna, M. (2002), 'Thinking and Caring about Cognitive Inconsistency: When and for Whom Does Attitudinal Ambivalence Feel Uncomfortable?', *Journal of Personality and Social Psychology* 82/2: 157–166.

Ochsner, K., and Gross, J. (2005), 'The Cognitive Control of Emotion', *Trends in Cognitive Sciences* 9/5: 242–249.

Payne, B., Vuletich, H., and Lundberg, K. (2017), 'The Bias of Crowds: How Implicit Bias Bridges Personal and Systemic Prejudice', *Psychological Inquiry* 28/4: 233–248.

Quilty-Dunn, J., and Mandelbaum, E. (2018a), 'Against Dispositionalism: Belief in Cognitive Science', *Philosophical Studies* 175/9: 2353–2372.

Quilty-Dunn, J., and Mandelbaum, E. (2018b), 'Inferential Transitions', *Australasian Journal of Philosophy* 96/3: 532–547.

Quine, W. (1951), 'Two Dogmas of Empiricism', *The Philosophical Review* 60/1: 20–43.

Quine, W., and Ullian, J. (1978), *The Web of Belief* (Random House).

Radvansky, G., and Copeland, D. (2006), 'Walking Through Doorways Causes Forgetting: Situation Models and Experienced Space', *Memory & Cognition* 34/5: 1150–1156.

Radvansky, G., Krawietz, S., and Tamplin, A. (2011), 'Walking Through Doorways Causes Forgetting: Further Explorations', *Quarterly Journal of Experimental Psychology* 64/8: 1632–1645.

Railton, P. (2014), 'Reliance, Trust, and Belief', *Inquiry* 57/1: 122–150.

Rauhut, A., Thomas, B., and Ayres, J. (2001), 'Treatments That Weaken Pavlovian Conditioned Fear and Thwart Its Renewal in Rats: Implications for Treating Human Phobias', *Journal of Experimental Psychology: Animal Behavior Processes* 27/2: 99–114.

Rayo, A. (2013), *The Construction of Logical Space* (Oxford University Press).

Rhodewalt, F., and Agustsdottir, S. (1986), 'Effects of Self-Presentation on the Phenomenal Self', *Journal of Personality and Social Psychology* 50/1: 47–55.

Ripley, D. (2009), 'Contradictions at the Borders', in R. Nouwen et al. (eds.), *International Workshop on Vagueness in Communication* (Springer), 169–188.

Ripley, D. (2011), 'Negation, Denial, and Rejection', *Philosophy Compass* 6/9: 622–629.

Rosch, E. (1978), 'Principles of Categorization', in B. Rosch and B. Lloyd (eds.), *Cognition and Categorization* (Erlbaum), 28–49.

Ross, M. (1989), 'Relation of Implicit Theories to the Construction of Personal Histories', *Psychological Review* 96(2): 341–357.

Sanborn, A., Griffiths, T., and Navarro, D. (2006), 'A More Rational Model of Categorization', in R. Sun and N. Miyake (eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 726–731.

Schul, Y., Mayo, R., and Burnstein, E. (2004), 'Encoding Under Trust and Distrust: The Spontaneous Activation of Incongruent Cognitions', *Journal of Personality and Social Psychology* 86/5: 668–679.

Schultheis, H., Barkowsky, T., and Bertel, S. (2006), 'LTM C: An Improved Long-Term Memory for Cognitive Architectures', *Proceedings of the Seventh International Conference on Cognitive Modeling*, 274–279.

Schwitzgebel, E. (2013), 'A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box', in N. Nottelmann (ed.), *New Essays on Belief* (Palgrave Macmillan), 75–99.

Seifert, C. (2002), 'The Continued Influence of Misinformation in Memory: What Makes a Correction Effective?', in B. Ross (ed.), *Psychology of Learning and Motivation*, vol. 41 (Academic Press), 265–292.

Shin, Y. S., Masís-Obando, R., Keshavarzian, N., Dáve, R., and Norman, K. A. (2020), 'Context-Dependent Memory Effects in Two Immersive Virtual Reality Environments: On Mars and Underwater', *Psychonomic Bulletin & Review* 1–9.

Sinnott-Armstrong, W. (2007), 'Abstract + Concrete = Paradox', in J. Knobe and S. Nichols (eds.), *Experimental Philosophy* (Oxford University Press), 209–230.

Smith, S., and Vela, E. (2001), 'Environmental Context-Dependent Memory: A Review and Meta-analysis', *Psychonomic Bulletin & Review* 8/2: 203–220.

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

Steele, C., Southwick, L., and Critchlow, B. (1981), 'Dissonance and Alcohol: Drinking Your Troubles Away', *Journal of Personality and Social Psychology* 41/5: 831–846.

Surowiecki, J. (2005), *The Wisdom of Crowds* (Anchor).

Tesser, A., and Cornell, D. (1991), 'On the Confluence of Self Processes', *Journal of Experimental Social Psychology* 27/6: 501–526.

Thibodeau, R., and Aronson, E. (1992), 'Taking a Closer Look: Reasserting the Role of the Self-Concept in Dissonance Theory', *Personality and Social Psychology Bulletin* 18/5: 591–602.

Thomson, D., and Tulving, E. (1970), 'Associative Encoding and Retrieval: Weak and Strong Cues', *Journal of Experimental Psychology* 86/2: 255–262.

Unkelbach, C. (2007), 'Reversing the Truth Effect: Learning the Interpretation of Processing Fluency in Judgments of Truth', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33/1: 219–230.

Vul, E., and Pashler, H. (2008), 'Measuring the Crowd Within: Probabilistic Representations within Individuals', *Psychological Science* 19/7: 645–647.

Wegner, D., Coulton, G., and Wenzlaff, R. (1985), 'The Transparency of Denial: Briefing in the Debriefing Paradigm', *Journal of Personality and Social Psychology* 49/2: 338–346.

Yalcin, S. (2018), 'Belief as Question-Sensitive', *Philosophy and Phenomenological Research* 97/1: 23–47.