The Fragmented Mind:
An Introduction

Dirk Kindermann & Andrea Onofri

This book is about the hypothesis that the mind is fragmented, or compartmentalized. When it is said that an agent's mind is fragmented, it is usually meant that their overall belief state is fragmented.[1] To a first approximation, a belief state can be said to be fragmented if it is divided into several sub-states of the same kind: fragments. Each fragment is consistent and closed under entailment, but the fragments taken together need not make for a consistent and closed overall state. The thesis that the mind is fragmented contrasts with the widespread, if often implicit, assumption—call it *Unity*—that the mind is unified, i.e. that an agent's overall belief state is consistent and closed under entailment. The motivation for fragmentation comes from a number of places, notably the shortcomings of *Unity*: the problem of logical omniscience, the problem of inconsistent doxastic states, cases of cognitive dissonance and imperfect information access, and others. In this introduction, we outline what varieties of fragmentation have in common and what motivates them. We then discuss the relationship between fragmentation and theses about cognitive architecture, introduce two classical theories of fragmentation, and sketch recent developments. Finally, as an overview of the volume, we present some of the open questions about and issues with fragmentation that the contributions to this volume address.

1. *Unity* and *Fragmentation*

Doxastic states like belief and epistemic states like knowledge are standardly assumed to be inherently rational. Much of epistemic logic, Bayesian accounts of human belief, decision theory, and some influential views about rationality proceed from the underlying view that the mind—or at least doxastic states—is unified:

Unity

Agents have a unified representation of the world (at time *t*)—a single state of belief organized by two principles:

1. *Consistency*: The total set of an agent's beliefs (at *t*) is consistent.

[1] This formulation focuses on belief, since this has been the main focus of the literature on the topic. We mean to leave open the possibility that a fragmented mind is one in which overall attitude states like knowledge or desire are (also) fragmented.

2. *Closure*: The total set of an agent's beliefs (at *t*) is logically closed. That is, agents believe the logical consequences of their beliefs.[2]

Unity may seem to impose unreasonably strong requirements on doxastic states. In the literature, *Unity*—as we call it—is often endorsed with one of the following qualifications.

First, *Unity* is often thought of as part of a descriptive theory of *ideal* rational agents, not of real agents. Thus, some authors implicitly or explicitly take their analyses to describe suitably idealized versions of real agents.[3] Idealization affords many theoretical advantages, including simplicity in accounting for logical relations among beliefs. None of this entails that the theory applies directly to real agents. An open question here is of course whether real agents are similar enough to these idealized agents for the theory to have any use in the explanation of real agents' doxastic attitudes (see Section 2.3 below).

Second, the *Consistency* and/or *Closure* principles are sometimes weakened. For instance, Easwaran and Fitelson (2015) propose "non-dominance" for an (ideally rational) agent's overall beliefs, a coherence requirement that is strictly weaker than *Consistency*.[4] *Closure* is sometimes weakened to apply only to logical consequences that would be "*manifest* to an ideal cognizer" (Fine 2007: 48; our italics) or to consequences the agent also *believes* to be consequences of her beliefs. Despite variance in the principles, most authors assume that they apply to an agent's single overall belief state.

Third, *Consistency* and *Closure* are sometimes understood as *rationality conditions* on belief. They are meant to be part of a normative account of how real agents *should* behave in their doxastic lives to count as rational.[5] It is sometimes not entirely clear whether a given analysis of belief is proposed as a *descriptive* account of real agents or as a *normative* account of what an agent's rational beliefs should be like. But even when a normative or heavily idealized account is assumed, most authors in philosophy, formal epistemology, and decision theory presuppose that agents have a single belief state.

Because of its theoretical virtues, *Unity* is widely adopted in several different areas. First, in Bayesian epistemology, in doxastic and epistemic logic, and in decision theory, *Unity* allows for much simpler formal models of human thought and agency. Thus, giving up on *Consistency* or

[2] To be inclusive, we'll keep the notions of consistency and closure quite general. A set of full beliefs thus counts as consistent just in case it contains no two beliefs whose contents cannot be true together. We will, for the most part, talk about full belief; where a graded notion of belief is relevant, consistency is understood as follows: A set of graded beliefs is probabilistically consistent *iff* it obeys the laws of probability. A belief counts as a logical consequence of a set of beliefs in case it follows from the set by the laws of logic.

[3] Stalnaker (1991) gives a useful overview of the role of idealization in theories of belief. See Griffith et al. (2012) on rationality as a methodological assumption in the descriptive analysis of human belief.

[4] Others, however, assume a notion of coherence that is stronger than *Consistency*: Traditional coherence theorists in epistemology require (rational, or justified) beliefs to be logically consistent *and* mutually inferable *and* to stand in various explanatory relations to each other (see e.g. BonJour 1988).

[5] See, for instance, Kolodny (2008) for discussion of descriptive vs. normative versions of *Consistency* and *Closure*.

Closure comes at the cost of giving up classical logical assumptions that underlie much of Bayesian probability theory and decision theory and would require the adoption of weaker, non-normal doxastic and epistemic logics.[6]

Second, in philosophy, most authors accept *Consistency* as a minimal requirement on rational belief (see Easwaran and Fitelson 2015). Kolodny's (2008) use of the term "the myth of formal coherence" for a package of principles including *Consistency* and *Closure* bespeaks their pervasiveness in the discipline.[7] For many authors, part of what it is to be a belief is to tend to produce beliefs in what follows from the belief in question and to tend to eliminate inconsistencies.[8] For instance, the Davidsonian tradition holds that, in order to interpret other people in linguistic communication, we must presuppose that their beliefs are consistent to a sufficient degree.[9] And Dennett (1981) explicitly claims that, when attributing intentional states to others in order to explain their behavior, our "starting" assumption is that their beliefs are consistent and deductively closed—we then revise that assumption on the basis of the specific circumstances of the agent we are dealing with.

Indeed, the assumptions of *Consistency* and *Closure* are reflected in our everyday, common-sense attributions of belief. We usually expect other agents to exhibit relevant reasoning patterns on the basis of their beliefs, drawing various kinds of inferences both consciously and unconsciously. We also expect that, upon receiving new information that contradicts their previously held beliefs, they will usually update their doxastic state accordingly, instead of simply accepting blatantly inconsistent propositions. Everyday belief attributions thus manifest both an assumption that people's beliefs are consistent and deductively closed and an expectation that people's beliefs should exhibit these two properties.

Finally, *Unity*'s plausibility stems in part from the idea that the point of belief is to represent the world accurately, without missing important bits.[10] But our beliefs, so this line of thought goes, cannot represent accurately and comprehensively if they are inconsistent and their consequences are not drawn.[11] In sum, according to this view of belief,

> beliefs … necessarily, or constitutively, tend to formal coherence as such (even if this tendency is sometimes inhibited). Part of what it is to be a belief, many in the philosophy of mind will say, is to tend to produce beliefs in what follows … And part of what it is to

[6] See, e.g., Christensen (2004: ch. 2) for discussion of the foundations of probabilistic analyses of belief in classical deductive logic. See Kaplan (1996) for a defense of *Consistency* and *Closure* at the heart of decision theory.

[7] Among the most outspoken supporters of *Consistency* are coherence theorists of epistemic justification (e.g. BonJour 1988). For recent attacks on *Consistency* which nonetheless presume that agents have a single total set of beliefs, see Kolodny (2007, 2008) and Christensen (2004).

[8] See, for instance, Bratman (1987), who holds this view for other attitudes, such as intentions.

[9] See Davidson (1973, 1982/2004).

[10] See, e.g., van Fraassen (1995: 349).

[11] The argument can be found, e.g., in Lehrer (1974: 203). Christensen (2004: ch. 4) gives a helpful summary of arguments for *Consistency*.

> be a belief is to tend either to repel contradictory beliefs, or to give way to them, as such. (Kolodny 2008: 438)

Despite *Unity*'s attractive features, applying *Consistency* and *Closure* to an agent's global doxastic state also generates serious problems, which we will discuss in the next section. Fragmentation views were originally proposed to avoid such problems while keeping with the idea that, locally, *Consistency* and *Closure* are important ingredients of what it is to be in a doxastic state such as belief. Davies and Egan provide the following summary of what it is for an agent to be in a fragmented belief state:

> Actual belief systems are fragmented or compartmentalised. Individual fragments are consistent and coherent but fragments are not consistent or coherent with each other and different fragments guide action in different contexts. We hold inconsistent beliefs and act in some contexts on the basis of the belief that P and in other contexts on the basis of the belief that not-P. Frequently we fail to put things together or to "join up the dots." It can happen that some actions are guided by a belief that P and other actions are guided by a belief that if P then Q, but no actions are guided by a belief that Q because the belief that P and the belief that if P then Q are in separate fragments. (Davies and Egan 2013: 705)

At a high level of abstraction, the hypothesis of fragmentation, or compartmentalization, consists of four claims:

Fragmentation

F1. The total set of an agent's beliefs (at time *t*) is fragmented into separate belief states.
F2. Each belief state (at *t*) is a fragment whose constituent beliefs are consistent with each other and closed under logical consequence.
F3. The belief states of a single agent (at *t*) are logically independent: They may not be consistent with each other, and the agent may not believe the consequences of his belief fragments taken together.
F4. Different belief fragments of a single agent (at *t*) guide the agent's actions in different contexts or situations.

Fragmentation sees an agent's overall belief state as fragmented into various sub-states, each of which is "active" or "available" for guiding action in a specific context. While each single sub-state, or fragment, is consistent and closed under logical consequence, an agent may have an inconsistent overall belief state on account of inconsistent beliefs belonging to different fragments, that is, inconsistent beliefs that are "active" in different contexts; furthermore, the agent may fail to draw the logical consequences of beliefs belonging to different fragments. (Note that theses F1–F4 do not entail that a single belief must belong to only one fragment; in fact, many (basic) beliefs will belong to most fragments if they are action-guiding in many contexts.) In what follows, we treat views that endorse fragmentation as minimally committed to claims F1–F4 (or similar versions thereof).

2. Motivations for fragmentation

2.1 Logical omniscience and closure

Some of the motivation for *Fragmentation* stems from the notorious shortcomings of *Unity*. For instance, *Unity*'s *Closure* principle faces a version of the problem of logical omniscience.[12] Let's consider the following, single-premise logical closure principle:

> **Single-premise closure**
> For any propositions *p* and *q*, if an agent *A* believes *p*, and *p* logically entails *q,* it follows that *A* also believes *q*.

Agents automatically believe all of the logical consequences of any of their beliefs. It is obvious that a view endorsing *Closure* is not descriptively adequate for agents with finite logical abilities (Parikh 1987, 1995); it implies that humans have no need for logical reasoning, as they already know all the logical consequences of any of their beliefs. But as Harman has famously argued, endorsing *Closure* is not normatively adequate either: Rational agents with "limited storage capacity" should not strive to draw any and all logical consequences from the beliefs they hold, or else their finite minds will become "cluttered" with trivial beliefs that are irrelevant to their lives (Harman 1986: 13).

One way for a theory of belief to avoid the problem of logical omniscience is simply by dropping any logical constraints on the overall set of an agent's beliefs. But as many have argued, a belief set must meet some minimal logical standards in order to count as a belief set at all. Thus, Cherniak writes,

> [t]he elements of a mind—and, in particular, a cognitive system—must *fit together* or cohere. A collection of mynah bird utterances or snippets of the *New York Times* are chaos, and so (at most) just a sentence set, not a belief set. ... no rationality, no agent. (Cherniak 1986: 6)

[12] The problem of logical omniscience is often associated with possible worlds models of attitude content. Note, however, that any view on which *Closure* is endorsed will face the problem, no matter what model of attitude content is adopted (cf. Stalnaker 1991 and Greco, Chapter * in this volume).

The problem of logical omniscience is thus but one side of the problem of finding the "right" consequence relation (or even more broadly, the "right" logical principles) under which the belief sets of real agents are closed (cf. Stalnaker 1991).

Fragmentation views of belief promise to make *some* progress on this larger problem by avoiding the counterintuitive results of multiple-premise closure:

> **Multiple-premise closure**
> For any propositions {*p*, *q*, … *r*}, if an agent *A* believes *p* and *A* believes *q* … and {*p*, *q*, …} together entail *r*, then *A* believes *r*.

Versions of the fragmentation approach deny that multiple-premise closure holds for the entire belief state, that is, across fragments. So if the belief that *p* and the belief that *q* belong to different fragments, the agent need not count as believing any logical consequence of *p* and *q* taken together. In addition, fragmentation views allow us to count such agents as rational, although there is a largely open question as to what a minimal threshold for rational belief across fragments might be on fragmentation views (cf. Cherniak 1986 and Borgoni, Chapter * in this volume). At the same time, fragmentation views do impose minimal constraints on belief sets, or states, by accepting versions of *Consistency* and *Closure* for the individual fragments. Thus, beliefs are "locally" consistent and complete, but not across fragments.[13]

2.2. Lewis on inconsistent beliefs

A parallel problem arises for *Unity*'s *Consistency* principle: It seems clear that agents sometimes hold inconsistent beliefs. Lewis (1982) provides a famous example:

> I used to think that Nassau Street ran roughly east–west; that the railroad nearby ran roughly north–south; and that the two were roughly parallel. (By "roughly" I mean "to within 20 degrees.") (Lewis 1982: 436)

Lewis's three beliefs were manifestly inconsistent. However, according to the *Consistency* principle that is part of *Unity*, agents do not hold inconsistent beliefs. Lewis's example thus seems to present a problem for *Unity*.

We might be tempted to simply dismiss *Consistency* as being implausibly strong and clearly inadequate. As noted in the previous section, however, this kind of move is not enough to solve the problem faced by *Unity*. We would not be inclined to consider a large collection of completely

[13] It is still part of fragmentation views that single-premise closure holds within every single fragment, and thus they still face the problem from single-premise closure. See Yalcin (2008, 2018) for an attempt to make progress on the single-premise closure problem for fragmentation views—cf. Section 3.4 of this introduction for more details. Rayo (2013) addresses the problem of mathematical and logical omniscience on a possible worlds conception of content.

inconsistent beliefs an agent’s doxastic state, so if we abandon *Consistency*, we will need an alternative principle: An agent’s beliefs might occasionally be inconsistent, as in Lewis’s case, but they cannot be systematically inconsistent while still counting as a cognitive system. So what principle should we adopt when dealing with the possibility of inconsistent beliefs?

Again, *Fragmentation* can be used to offer an answer to this question. Immediately after sketching the above case, Lewis himself provides a brief but influential formulation of the fragmentation approach:

> Now, what about the blatantly inconsistent conjunction of the three sentences? I say that it was not true according to my beliefs. My system of beliefs was broken into (overlapping) fragments. Different fragments came into action in different situations, and the whole system of beliefs never manifested itself all at once. The first and second sentences in the inconsistent triple belonged to—were true according to—different fragments; the third belonged to both. The inconsistent conjunction of all three did not belong to, was in no way implied by, and was not true according to, any one fragment. That is why it was not true according to my system of beliefs taken as a whole. Once the fragmentation was healed, straightway my beliefs changed: now I think that Nassau Street and the rail-road both run roughly northeast–southwest. (ibid.)

Lewis’s system of beliefs consists of (is “broken into”) different fragments, none of which includes the inconsistent conjunction of all three beliefs. Therefore, each fragment respects consistency requirements—the inconsistency would only arise in a fragment that included all three beliefs, but there is no such fragment in Lewis’s overall doxastic state. Within a fragmentation approach, then, individual fragments are internally consistent, but inconsistency may still arise among beliefs belonging to different fragments (as in Lewis’s case).[14]

2.3 Idealization and explanatory power

In belief–desire psychology, doxastic states are assumed to figure in psychological laws that explain and predict the actions of rational agents, such as “[I]f A wants p and believes that doing q will bring about p, then ceteris paribus, A will do q” (Borg 2007: 6).[15] Arguably, such laws make use of our everyday notions of belief and desire, if not of scientifically respectable ones. The endorsement of this role of belief in the explanation and prediction of action, however, presents a difficulty for views that (even tacitly) endorse *Unity*. *Unity* is descriptively inaccurate; finite agents are not always consistent in their beliefs and do not draw each and every deductive inference from them. As a result, the assumptions of *Consistency* and *Closure* typically involve

[14] For discussion of Lewis’s case, see for instance the contributions to this volume by Borgoni, Egan, Greco, and Yalcin.

[15] Cf. Davidson (1963, 2004).

the idealization of agents' rational capacities. While idealization is a common and useful element of scientific theorizing, the assumption of ideal rationality—exemplified in *Consistency* and *Closure*—seems to be so extreme that it cannot be applied in interesting ways to real agents with finite computational capacities and limited memory. In particular, ideal rationality appears to leave those theories without any ability to explain or predict the actions of real agents at all.[16] Cherniak (1986) uses the following story to make his case against the assumption of agents with idealized rational capacities:

> In "A Scandal in Bohemia" Sherlock Holmes's opponent has hidden a very important photograph in a room, and Holmes wants to find out where it is. Holmes has Watson throw a smoke bomb into the room and yell "Fire!" when Holmes's opponent is in the next room, while Holmes watches. Then, as one would expect, the opponent runs into the room and takes the photograph from its hiding place. ... once the conditions are described, it seems very easy to predict the opponent's actions. Prima facie, we predict the actions ... by assuming that the opponent possesses a large set of beliefs and desires—including the desire to preserve the photograph, and the belief that where there's smoke there's fire, the belief that fire will destroy the photograph, and so on—and that the opponent will act appropriately for those beliefs and desires. (Cherniak 1986: 3–4)

Against theories of belief that assume the idealized rationality of belief, desire, and their connection to action, Cherniak argues that

> with only such a theory, Holmes could not have predicted his inevitably suboptimal opponent's behavior on the basis of an attribution of a belief–desire set; he could not have expected that her performance would fall short of rational perfection in any way, much less in any *particular* ways. Holmes would have to regard his opponent as not having a cognitive system. (Cherniak 1986: 7)

The idealizations involved in *Unity*, then, make the theory virtually inapplicable to agents "in the finitary predicament" (Cherniak 1986: 8). Since we, like Holmes, do have ways of explaining and predicting agents' specific behavior based on attributions of beliefs and desires, no such theory can be true of our notion of belief. Put differently, if we want a theory of belief that allows for the explanation and the prediction of the actions of real agents, we'd better let go of the extreme idealizations involved in *Unity*.

[16] We are adapting Cherniak's argument against what he calls the "ideal general rationality condition" ("If A has a particular belief–desire set, A would undertake *all* and only actions that are apparently appropriate," Cherniak 1986: 7). See also Stalnaker (1991) for arguments for why the standard reasons for idealizing do not apply to the idealizations underlying *Consistency* and *Closure*.

At the same time, we saw in the previous sections that we cannot let go of all rationality constraints on the notion of belief either. If our theory allowed for agents' overall belief states to be random sets of beliefs without any restrictions, it is not clear that we'd have a theory of the beliefs of *agents.* Cherniak concludes that we need to steer a middle path between idealized rationality conditions and no rationality conditions—what he calls *minimal rationality*. Fragmentation views fit minimal rationality constraints: Agents do not maintain consistency or logical closure across their entire belief set, or state, and neither do they need to in order to count as minimally rational. Fragmentation views allow for inconsistency and closure failure across fragments. But agents do (need to) keep their beliefs locally consistent and logically closed. Fragmentation views do place consistency and closure requirements on each individual fragment.

2.4 Memory and information access

Another motivation in favor of fragmentation starts with the observation that a piece of information we possess (in memory) may be accessible to us given one purpose or task but may not be accessible given another purpose or task. Consider the following example offered by Stalnaker:

> [I]t will take you much longer to answer the question, "What are the prime factors of 1591?", than it will the question, "Is it the case that 43 and 37 are the prime factors of 1591?" But the answers to the two questions have the same content, even on a very fine-grained notion of content. Suppose that we fix the threshold of accessibility so that the information that 43 and 37 are the prime factors of 1591 is accessible in response to the second question, but not accessible in response to the first. Do you know what the prime factors of 1591 are or not? (Stalnaker 1991: 438)

The less mathematically inclined among us are perhaps just as Stalnaker describes: The information that *43 and 37 are the prime factors of 1591* is accessible to us for the task of answering the yes or no question, but it is not accessible for the task of answering the *wh*-question.[17] Stalnaker (1991, 1999) argues that a notion of accessible belief (relative to a purpose) is necessary in explanations of action in terms of belief and desire. Some of an agent's actions are best predicted or explained by attributing a belief *p* to them (given their desires), while other actions performed by the same agent in the same situation can only be predicted or explained when *p* is not among their beliefs (given the same desires). The notion of an accessible, or available, belief can be naturally accommodated by the fragmentation approach. For instance, variable information access (in memory) may be due to the need for an efficient recall mechanism

[17] Elga and Rayo (Chapter * in this volume) and Rayo (2013) provide a number of further examples that support the claim that information is accessible to us relative to various purposes, or tasks. See also Greco (2019) for the claim that an agent's knowledge and evidence are available only relative to particular purposes.

that searches only those parts of memory that are related to the purpose at hand (Cherniak 1986: ch. 3). This assumption sits naturally with the view according to which our information (in memory) is organized into fragments, which in turn are associated with particular purposes, or tasks.

2.5 The preface paradox

Next, fragmentation may allow for an intuitive solution to the preface paradox. It is rational for a book's author to believe all the statements they make in the book. At the same time, it is rational for them to admit their fallibility, that is, to state *F*: that at least one of these beliefs is false. Thus, it would be rational for them to have inconsistent beliefs: a set of beliefs *and* the belief that at least one of them is false.

On fragmentation views, Cherniak argues, it may indeed be (minimally) rational to have inconsistent beliefs in genuine preface paradox cases:

> The seemingly overlooked point that is of interest here is that the *size* of the belief set for which a person makes the statement of error *F* determines the reasonability of his joint assertions. If he says, "Some sentence in {*p*} is false, and *p*," this seems clearly irrational, like saying, "I am inconsistent; I believe both *p* and not-*p.*" If he says, "Some sentence in {*p*, *q*} is false, and *p*, and *q*," this is similarly unacceptable. But if the set is very large, and in particular encompasses the person's total belief set, then accepting *F* along with that belief set becomes much more reasonable. (Cherniak 1986: 51)

Within a single fragment, some version of *Closure* holds from which the principle of *Agglomeration* follows for beliefs *p* and *q* in the same fragment: ($B(p)$ & $B(q)$) → B(*p* & *q*), where *B* is the belief operator. Taking *Closure* as a constraint on rational belief, a fragmentation theorist may hold that it is irrational to add to a fragment a belief *F* in one's fallibility with regard to the beliefs in that fragment. Otherwise, there would be an inconsistent fragment. But for larger sets of beliefs including *p, q, r, ...* that belong to different fragments, it is not irrational to add to one's set of beliefs the belief *F* that at least one of *p, q, r, ...* is false. No closure condition holds across fragments, so *Agglomeration* doesn't hold either for beliefs *p, q, r, ...* and *F*, and no inconsistent conjunction need be derived.[18]

2.6 Cognitive dissonance and implicit bias

Cases of "cognitive dissonance" and "implicit bias" seem to provide further support for fragmentation. Before we move to the cases, it will be helpful to offer a general characterization

[18] Stalnaker's fragmentation approach to the preface paradox is somewhat different from Cherniak's position above. While Stalnaker (1984: ch. 5) agrees that *Agglomeration* should be given up as a global, i.e. inter-fragment, normative ideal of rationality for the attitude of "acceptance," he upholds *Agglomeration*, together with *Consistency*, as a global normative ideal of rationality for the attitude of belief.

of these two related phenomena, starting with cognitive dissonance. Aronson (1997) describes cognitive dissonance as follows, as he summarizes the main ideas behind Festinger's classic work on the topic:

> If a person holds two cognitions that are psychologically inconsistent, he experiences *dissonance*: a negative drive state (not unlike hunger or thirst). Because the experience of dissonance is unpleasant, the person will strive to reduce it—usually by st[r]uggling to find a way to change one or both cognitions to make them more consonant with one another. (Aronson 1997: 128)

As an example of two dissonant psychological states, consider the following case by Gendler (2008a):

> Last month, when I was traveling to the APA Program Committee meeting, I accidentally left my wallet at home. … when I got to Baltimore, I arranged to borrow money from a friend who was also attending the meeting. As he handed me the bills, I said: "Thanks so much for helping me out like this. It is really important for me to have this much cash since I don't have my wallet." Rooting through my bag as I talked, I continued: "It's a lot of cash to be carrying loose, though, so let me just stash it in my wallet ..." (Gendler 2008a: 637)

Gendler herself provides a concise statement of the problem raised by this case:

> How should we describe my mental state as my fingers searched for my wallet to house the explicitly wallet-compensatory money? (Gendler 2008a: 637)

Gendler clearly seems to believe that she does not have her wallet—indeed, she explicitly says so to her friend. At the same time, however, her behavior does not fit with the belief she has openly expressed. Given her belief, her non-verbal behavior (moving her fingers to search for the wallet) will yield no useful outcome, for she knows the wallet is simply not there. The same applies to her verbal behavior ("let me just stash it in my wallet"); given her belief that the wallet is not there, the intention that she is verbally expressing cannot be fulfilled.

To see why cognitive dissonance is a problem, note that belief is generally taken to have an intimate connection with action. This seems to be part of our "folk," commonsense conception of belief—put simply, we expect people to act in ways that fit well with their beliefs, and when they do not act in those ways, we start doubting whether they really do have the beliefs in question. Furthermore, the belief–action connection is central to many philosophical accounts of belief (see for instance Stalnaker 1984 and Greco, Chapter * in this volume). Once this idea is in place, it is easy to see why cognitive dissonance is a problem. Part of Gendler's verbal behavior ("I don't have my wallet") indicates that she believes: *I don't have my wallet.* However, other aspects of her behavior (such as searching for the wallet) indicate that she believes: *I have my wallet*. Now, if we adopted *Unity*'s requirement of *Consistency*, one of these options must be incorrect, for the two beliefs are of course inconsistent.

Fragmentation offers a more promising solution. Gendler's initial assertion ("I don't have my wallet") is guided by information in one fragment, a fragment which includes the belief *I don't have my wallet*. Gendler's other actions (such as searching for the wallet) are guided by information in a different fragment, a fragment which includes the belief *I have my wallet.* Gendler holds both beliefs, so we can make sense of her behavior; however, the two beliefs are stored in different fragments, so intra-fragment consistency is preserved.

Let us now move on to "implicit bias." Brownstein (2019) defines the notion and provides an example:

> "Implicit bias" is a term of art referring to relatively unconscious and relatively automatic features of prejudiced judgment and social behavior. … the most striking and well-known research has focused on implicit attitudes toward members of socially stigmatized groups, such as African-Americans, women, and the LGBTQ community … For example, imagine Frank, who explicitly believes that women and men are equally suited for careers outside the home. Despite his explicitly egalitarian belief, Frank might nevertheless behave in any number of biased ways, from distrusting feedback from female co-workers to hiring equally qualified men over women. (Brownstein 2019)

Brownstein offers several empirically documented examples of implicit bias, but Frank's imaginary case will be enough for our purposes.[19] The central question is: What does Frank believe concerning gender equality? His explicit verbal behavior suggests that he has the following belief: *Women and men are equally suited to careers outside the home*. At the same time, other aspects of his behavior (like his behavior in the workplace) suggest that he believes that *women and men are not equally suited to careers outside the home.*

Again, since the two beliefs are inconsistent, Frank cannot have both of them, according to unity. Fragmentation offers an alternative.[20] Frank's overall doxastic state could be divided into various fragments, with one fragment including his egalitarian belief and guiding his explicit verbal behavior and another fragment including his anti-egalitarian belief and guiding other aspects of his behavior, such as his decisions in the workplace. As in Gendler's cognitive

[19] See Schwitzgebel (2010) for a more detailed example of the same kind.

[20] For fragmentation approaches to cognitive dissonance and implicit bias, see for instance Bendana (Chapter * in this volume), Borgoni (2018), and Mandelbaum (2015).

dissonance case, then, fragmentation might give us new explanatory tools to make sense of "inconsistent" behavior exhibited by implicitly biased subjects like Frank.[21]

3. Fragmentation and cognitive architecture

3.1 Horizontal fragmentation

It's useful to distinguish fragmentation views from a cluster of views which have also received the label "fragmentation" (cf. Greco 2014). We can call fragmentation as we understand it here "horizontal" (or intra-attitude-type) fragmentation as it involves the claim that the *same kind* of attitude or representational state is fragmented; that is, an agent's *beliefs* are said to be divided into different fragments.

A contrasting family of views have been offered as explanations of cognitive dissonance (see Section 2.6). Despite great differences in detail, these views share the idea that different *kinds* of representational mental states are involved in, and responsible for, the manifestations of our apparent beliefs: assertions and explicit avowals, on the one hand, and more automatic, action-entrenched responses, on the other hand. One prominent version of this idea is developed by Gendler (2008a, 2008b). Gendler draws a distinction between belief and what she calls "alief," where the former is responsible for explicit avowals and the latter is "*a*ssociative, *a*utomatic, and *a*rational" (Gendler 2008a: 641).

This family of views can be called "vertical" fragmentation (or inter-attitude-type fragmentation) as it posits different *kinds* of representational mental states. This is different from the "horizontal" (or intra-attitude-type) fragmentation discussed so far. While horizontal and vertical fragmentation views are in principle compatible, they present *prima facie* competing explanatory strategies regarding the phenomenon of cognitive dissonance.[22] Since horizontal fragmentation views will be our main focus, we will only use "fragmentation" to refer to this group of views in what follows.

3.2 Fragmentation, cognitive processes, modularity

Another question concerns the connection between fragmentation and cognitive architecture. By saying that an attitude like belief is fragmented *in the minimal sense of F1–F4 above*, one is

[21] There are other arguments against unity and in favor of fragmentation that we will not be able to discuss here. In particular, the hypothesis of fragmentation plays an important role in Davidson's (1982/2004, 1986/2004) account of rationality and action, in Egan's (2008) theory of perception and belief formation, and in Greco's (2014) discussion of epistemic akrasia. These potential applications of fragmentation are important and should be explored in detail, but here we prefer to focus on a narrower set of issues to avoid making the discussion too difficult to follow. The interested reader is also referred to Bendana and Mandelbaum (Chapter * in this volume) for a presentation of empirical evidence in favor of *Fragmentation*.

[22] According to our classification, "vertical fragmentation" views include, e.g., Bilgrami (2006), who distinguishes committal from dispositional beliefs; Gendler (2008a), who contrasts alief with belief; and Gertler (2011), who separates occurrent from dispositional beliefs.

not automatically making a claim about the psychological processes or systems that produce beliefs. An agent's overall belief state may be split into several fragments, but this does not imply that beliefs in the same fragment are all the result of the same type of psychological process, nor that beliefs in different fragments must be the result of different processes. For instance, theories of fragmentation are distinct from (but compatible with) dual-process and dual-system views.[23] Of course, this doesn't prevent fragmentation theorists from providing a substantial account of how fragmentation is implemented cognitively—see Bendana and Mandelbaum (Chapter * in this volume) for such a proposal.

For parallel reasons, we also take fragmentation to be *prima facie* distinct from modularity.[24] If fragmentation about belief turns out to be correct, then the beliefs of an ordinary cognitive agent belong to different fragments. However, it does not follow that some of those beliefs belong to different cognitive modules, for they might all be processed in a central, non-modular way. So the two notions (fragmentation and modularity) yield cross-cutting distinctions among cognitive states. Again, this is just to say that the two notions are conceptually distinct, not that there cannot be interesting connections between the phenomena in question.

In sum, to be a (horizontal) fragmentation view means being committed to (something like) claims F1–F4 in Section 1, and this commitment is compatible with different answers to questions about cognitive architecture, psychological processes, systems, and so forth. As we will see, many theories of fragmentation do take on more substantial commitments regarding the nature of belief and/or cognitive architecture. Still, these additional commitments are not immediate consequences of the core theses that define the fragmentation hypothesis.

4. Theories of fragmentation

The idea of fragmentation was powerfully expressed in philosophy by a few authors in the 1980s but seems to have lain dormant since then. It is only in recent years that the idea has attracted renewed interest. Here, we introduce two influential early theories of fragmentation, by Christopher Cherniak and by Robert Stalnaker, and sketch some of the recent developments.

4.1 Cherniak on memory and cognitive architecture

In his book *Minimal Rationality* (1986), Christopher Cherniak develops an account of rationality in which fragmentation plays an important role.[25] The main motivation behind Cherniak's theory is succinctly stated in the following passage:

[23] See Frankish (2010) for an overview of dual-process and dual-system views.

[24] See Fodor (1983).

[25] Cherniak generally uses the term "compartmentalization," but he does sometimes speak of "fragments" (see for instance Cherniak 1986: 68). Here we use the terms "compartmentalization" and "fragmentation" interchangeably.

> How rational must a creature be to be an agent, that is, to qualify as having a cognitive system of beliefs, desires, perceptions? Until recently, philosophy has uncritically accepted highly idealized conceptions of rationality. But cognition, computation, and information have costs; they do not just subsist in some immaterial effluvium. We are, after all, only human.
>
> … an agent can have a less than perfect deductive ability. I will argue that, although in everyday psychological explanations of behavior we require rationality of an agent, we in fact require only minimal, as distinguished from ideal, rationality. (Cherniak 1986: 3)

Cherniak then moves on to argue against highly idealized theories of rationality and develops his alternative theory of "minimal rationality." Chapter 3 of Cherniak's work is especially interesting for our purposes. Here, Cherniak focuses on memory—more specifically, he argues that traditional views of rationality have presupposed an idealized model of memory, one that imposes unrealistic demands on finite cognitive agents like us. He then proposes an alternative, more realistic model of memory in which fragmentation is central.

Cherniak spells out the problem faced by traditional, idealized models in his discussion of the following case:

> Smith believes an open flame can ignite gasoline (he uses matches to light bonfires, etc.), and Smith believes the match he now holds has an open flame (he would not touch the tip, etc.), and Smith is not suicidal. Yet Smith decides to see whether a gasoline tank is empty by looking inside while holding the match nearby for illumination. (Cherniak 1986: 57)

It seems that Smith has on this occasion failed "to infer the obvious conclusion that his match might ignite the gasoline" (Cherniak 1986: 57). But what would an account that modeled cognitive agents as ideally rational say about this case? According to such an account, Cherniak says,

> Smith cannot believe that the match has a flame because if he did, he must—by the ideal rationality condition that he make all useful inferences—conclude that holding it near the tank is dangerous; and he does not do this. (Cherniak 1986: 58)

However, Cherniak notes, it seems incorrect to hold that Smith does *not* believe that the match has a flame just because he has failed to draw a useful inference.[26] To offer an alternative

[26] Note that Smith's case is a clear example of an agent who fails to be logically omniscient—see Section 2.1 for discussion.

account, Cherniak draws on a classic distinction from cognitive psychology—that between short-term or active memory and long-term memory. Short-term memory has limited storage capacity and limited duration, but the agent can operate on its contents—items that are active in short-term memory can serve as premises for inferences, act as an input to practical reasoning, and so on. On the contrary, long-term memory has "practically no capacity limit" (Cherniak 1986: 53), but the subject cannot operate directly on items stored in it—she must first retrieve those items, so that they can be activated in short-term memory and be operated upon.

But retrieving items from long-term memory in a reliable and efficient way is a difficult cognitive task in itself. On the one hand, we want our searches to be reliable: For instance, we want to retrieve information that is relevant to the cognitive problem that we are currently trying to solve. On the other hand, we also want our searches to be efficient: We have limited time to conduct our search, so it is not possible to check all of our memory and only retrieve those items that are relevant to our current cognitive task. An exhaustive search might be possible given unlimited time, but it is not possible for finite agents like us. We thus have to accept a "trade-off of reliability for speed" (Cherniak 1986: 66), conducting searches that are not perfectly reliable but can be performed in a limited amount of time.

This, Cherniak says, is achieved by organizing our memory—which is where Cherniak's notion of "compartmentalization" comes in. Items in long-term memory "must be organized into subsets according to subject matter, where items within a subset are more likely to be relevant to each other than items from different subsets" (Cherniak 1986: 66). Depending on the cognitive problem at hand, the agent will search within subsets that are most likely to be relevant to the problem in question. Since memory is organized in this way, the search is not random. Searching in a random way would be a completely unreliable recall method, often selecting pieces of information that have nothing to do with the current task, whereas searching within an organized memory allows one to select clusters of information that "match" the task. At the same time, however, the search is not perfectly reliable—sometimes, highly relevant information is neglected. Indeed, this is what happens in Smith's case:

> [J]ust this type of failure to recall appropriate beliefs has occurred when Smith holds the match near the gasoline tank; given Smith's goal of self-preservation, his beliefs about whether a flame will ignite gasoline, and so on, are obviously relevant to the question of whether or not he should attempt this action. (Cherniak 1986: 61)

We can now see how the idea of fragmentation can be used to explain the case that proved problematic for highly idealized models of rationality:

> [T]he belief that a flame can ignite gasoline is filed under, roughly, "means of ignition"; the belief that the match he now holds has a flame is filed instead under "means of

> illumination.” The “illumination” category rather than the “ignition” category was checked because Smith decided he needed more light to see into the tank. The two crucial beliefs here … therefore were not both in short-term memory to be “put together”; but only if they were being thought about together could Smith make the connection and infer that there was danger. (Cherniak 1986: 57)

On this model of memory, the kind of mistake Smith makes is to be expected. Compartmentalizing our memory allows us to conduct fairly reliable searches in a limited amount of time, but it also leads to errors.

The defender of idealized models of rationality might now observe that Smith’s behavior in the above case can be described as irrational. Since that behavior is partly a result of the way in which Smith’s memory is organized, is compartmentalization itself a less-than-rational cognitive strategy? Not at all, Cherniak says:

> [A] basic precondition for our minimal rationality is efficient recall, which itself requires incomplete search, which in turn requires compartmentalization. Compartmentalization seems in this way a fundamental constraint on human knowledge representation. (Cherniak 1986: 69–70)

> [C]ompartmentalization is not just a regrettable failing of human beings, a departure from rationality *simpliciter*. Narrowly viewed, it leads to irrational actions, but overall, given our limitations (in particular, the slowness of exhaustive search), memory ought to be compartmentalized. Global rationality requires some local irrationality. (Cherniak 1986: 70)

Of course, this does not mean that all ways of organizing information in compartments are equally good. This, however, does not affect Cherniak’s main point: Fragmentation plays an essential role in a realistic account of memory; in turn, such an account will be a central component in a plausible theory of rationality.

4.2 Stalnaker and the possible worlds framework

Stalnaker (1984) developed fragmentation as a strategy to deal with versions of the problem of logical omniscience as they arise for possible worlds accounts of propositional attitudes and their contents. Stalnaker breaks the problem down into three conditions on sets of propositions—we here present a version for belief:[27]

[27] Stalnaker discusses the problem as it arises for the attitude of acceptance, of which beliefs are a subclass.

1. If *P* is a member of a set of [believed] propositions, and *P* entails *Q*, then *Q* is a member of that set.
2. If *P* and *Q* are each members of a set of [believed] propositions, then *P & Q* is a member of that set.
3. If *P* is a member of a set of [believed] propositions, then *not-P* is not a member of that set. (Stalnaker 1984: 82)

Note that conditions 1 and 3 are akin to *Unity*'s *Closure* and *Consistency* principles. So if the possible worlds account of attitudes is committed to 1–3, it accepts *Unity*. Let's see why 1–3 apply to the possible worlds account. Suppose an agent believes that Vienna is the capital of Austria. This attitude state is understood, or modeled, in terms of the set of worlds compatible with the belief: the set of all and only worlds in which Vienna is the capital of Austria. Now, Stalnaker understands not only single beliefs but also an agent's overall belief state in terms of the set of worlds that are compatible with the agent's belief state. The set of propositions believed thus contains exactly those propositions that are true in all these worlds. (For a possible worlds proposition to be true in a world is for that world to be a member of the set that is that proposition.)

Take the agent who believes that Vienna is the capital of Austria and suppose that their overall belief state is understood in terms of the set of worlds in which Vienna is the capital of Austria. Since any world in which Vienna is the capital of Austria is a world in which Vienna is a city in Austria, the agent would also count as believing that Vienna is a city in Austria. After all, their belief state as defined by the set of worlds in which Vienna is the capital of Austria is a belief state compatible with Vienna's being a city in Austria. So condition 1 is fulfilled by the possible worlds account of attitude states that understands the state in terms of the set of worlds compatible with it, and the set of propositions to which the agent has the attitude as containing those propositions that are true in all of those worlds.

Next, suppose the agent believes that Vienna is the capital of Austria and also believes that Austria is north of Italy. Then their belief state is given by the set of worlds in which both of these propositions are true: the set of worlds in which Vienna is the capital of Austria and Austria is north of Italy. Of course, the conjunctive proposition that Vienna is the capital of Austria and Austria is north of Italy is also true in those worlds. So the agent counts as believing the conjunctive proposition that Vienna is the capital of Austria and Austria is north of Italy. In this way, the possible worlds account of attitude states fulfills the second condition.

Finally, the agent who believes that Vienna is the capital of Austria and whose belief state is given by the set of worlds in which Vienna is the capital of Austria cannot count as also believing that Vienna is not the capital of Austria. That's because the proposition that Vienna is not the capital of Austria is not true in any of the worlds in the agent's belief state, all of which are Vienna-is-the-capital-of-Austria worlds. So it can’t be a member of the set of propositions the agent believes. Condition 3 is also met on the possible worlds account.

Stalnaker’s commitment to conditions 1–3 for the possible worlds account of attitudes, and hence the problems of logical omniscience, come from understanding an agent’s attitude state in terms of the set of worlds compatible with that state (“the possibilities that remain open for [the] agent in the ... state,” Stalnaker 1984: 81). Stalnaker’s motivation for this understanding of attitude states, on which these strong conditions are in-built, comes from his underlying pragmatic-dispositional account of propositional attitudes:

> Beliefs, according to the pragmatic picture, are conditional dispositions to act. A rational agent is, in general and by definition, disposed to act appropriately, where what is appropriate is defined relative to his beliefs and desires. To say that an agent believes that *P* is to say something like this: the actions that are appropriate for that agent—those he is disposed to perform—are those that will tend to serve his interests and desires in situations in which *P* is true. (Stalnaker 1984: 82)

So belief is defined in terms of dispositions to act, given one’s desires—in fact, given one’s desires and also one’s other beliefs. On the pragmatic-dispositional account it is necessary, Stalnaker argues, to understand any single belief against the backdrop of a belief state. Stalnaker provides the following example for illustration:

> Suppose I believe, as I do [in 1984], that someone will be elected President of the United States in 1988. One way in which that proposition could be realized is for *me* to be the one elected, but I know that that is not the way my belief will come true. For my actions to be appropriate, given that I have this belief, it is surely not required that I take account of that possibility, since it is excluded by my other beliefs. The actions that are appropriate for an agent who believes that *P* depend not only on what he wants but also on what else he believes. So it is necessary to define appropriateness relative to a total set of beliefs, or a belief state. And all that matters about such a belief state, as far as the appropriateness of actions or the agent’s dispositions to act are concerned, are the entailments of the belief state. (Stalnaker 1984: 82)

For an agent to believe *P* is for them to have a disposition to act in ways that bring about their desires in a world in which *P*—and also *Q* and *R* and *S*, given that those are the agent’s (relevant) further beliefs. Thus Stalnaker’s disposition to act in 1984, given his belief that someone will be elected US president in 1988, is not best understood as including dispositions to think about his campaign platform or who to make the next secretary of state, since the possibility of his being elected is ruled out by his other beliefs. We don’t have dispositions to act in desire-fulfilling ways under just any circumstances.

So how does fragmentation help with the problem, given the pragmatic-dispositional picture’s commitment to belief states? As Stalnaker says, a single agent can be in more than one belief state at the same time:

> A person may be disposed, in one kind of context, or with respect to one kind of action, to behave in ways that are correctly explained by one belief state, and at the same time be disposed in another kind of context or with respect to another kind of action to behave in ways that would be explained by a different belief state. ... the agent might, at the same time, be in two stable belief states, be in two different dispositional states which are displayed in different kinds of situations. (Stalnaker 1984: 83)

According to Stalnaker, conditions 2 and 3 need not hold for the agent’s overall beliefs, for their dispositions to act need not be explained against the background of the totality of their beliefs. Consider again Lewis, who had beliefs in the inconsistent triad: that Nassau Street in Princeton ran roughly east–west, that the railroad nearby ran roughly north–south, and that the two were roughly parallel (cf. Section 2.2). Lewis’s overall disposition to act might best be understood, given his story, by his different dispositions to act. On the one hand, he has dispositions including one to take streets that run perpendicular to Nassau Street when he intends to walk to the railway station; these dispositions are best explained by attributing the beliefs that Nassau Street runs roughly east–west and that it runs roughly parallel to the railroad. On the other hand, he has dispositions including one to take Nassau Street when he intends to take his bike for a ride to Trenton, roughly south/southeast of Princeton, without having to cross the railroad tracks to New York City in the north; these dispositions are best explained by his beliefs that the railroad runs roughly north–south and that it is parallel to Nassau Street. Lewis’s overall dispositions to act are thus best explained by attributing different belief states, or fragments, to him, each of which explains dispositions in different contexts.

What about condition 1, a version of single-premise closure, which still holds for single belief states? Here, Stalnaker bites the bullet: “[T]here is no basis, given the pragmatic account of belief, for defining the set of propositions believed, relative to a belief state, in a way that conflicts with the first deductive condition” (Stalnaker 1984: 82). Why is that? Consider Stalnaker’s belief that someone will be elected US president in 1988. If Stalnaker’s actions in his example are appropriate (given his desires and further beliefs)—that is, if they tend to serve his interests and desires in situations in which someone will be elected US president in 1988, then they will also tend to serve his interests and desires in situations in which the 1988 US presidential elections aren’t cancelled because of a global pandemic. That’s because every situation in which someone is elected is one where the elections aren’t cancelled because of a global pandemic. So, on the pragmatic-dispositional account, Stalnaker also counts as believing that the elections won't be cancelled because of a global pandemic.[28]

The fragmentation of an agent's overall beliefs into separate belief states, or fragments, allows Stalnaker to explain the informational gain of deductive reasoning as resulting from the "integration of ... separate belief states" (1984: 98). Consider an agent with one belief state in which *P* is true and another belief state in which *Q* is true, but no belief state in which both *P* and *Q* are true. While some proposition *R* may be a deductive consequence of propositions *P* and *Q* taken together, *R* is not a proposition that is true in any of the agent's belief states. Only once the agent integrates their *P*-belief state and their *Q*-belief state are they in a belief state in which *R* is true. Hence the integration of the two belief states, in deductive reasoning, results in the agent's gaining a new belief.[29]

4.3 Recent developments

In recent years, some authors have returned to the idea of fragmentation in a variety of philosophical contexts. Here we will only mention a few such proposals. Our goal is neither to provide an exhaustive overview nor to examine the details of the mentioned views, but rather to give a general idea of some recent developments.

Fragmentation in epistemology

Cases like those involving implicit bias (Section 2.6) not only raise questions about the nature of the mental states involved but also bring epistemological issues to the fore. As Daniel Greco (2015) notes, such cases put pressure on iteration principles such as—for knowledge—"If *S* knows that *P*, then *S* knows that *S* knows that *P*." Consider the case of the unwitting historian, who in fact has knowledge of English history from secondary school—she is able to reliably answer questions about William the Conqueror, Queen Elizabeth, and so on—but who has forgotten that she ever even took English history courses and insists that she knows nothing about English history. It's natural to take cases like this as evidence against iteration principles: For many propositions *p* about English history, the unwitting historian knows *p* but does not know that she knows *p*. But while it's plausible that our mental lives aren't fully transparent to us, giving up on iteration principles entirely comes at a cost. For instance, they allow for a straightforward explanation of "dubious," Moore's paradox–like assertions of the form "*P* but I don't know whether I know that *P*" (Greco 2015: 11 ff.).

[28] In his discussion of two counterexamples to the first deductive condition, Stalnaker (1984: 88–9) considers defining "active belief" by adding a condition to the entailment, so that only propositions that are entailed by the belief state and meet this further condition are also in the belief state. See Yalcin (2018) for a partial solution to single-premise closure problems for possible worlds fragmentation views.

[29] Rayo (2013) develops a Stalnaker-inspired fragmentation account of deduction in more detail—see Section 4.3.

Fragmentation, Greco argues, allows for an explanation of implicit bias cases and the case of the unwitting historian while also salvaging qualified versions of iteration principles. The unwitting historian has stored her knowledge of English history in a fragment she can access for the purposes of answering questions about English history but for no other purposes, while her second-order knowledge is stored in a fragment tied to other purposes.[30] So the apparent failure of the iteration principle for knowledge is due to inter-fragment opacity: Iteration principles hold within each single fragment—"If *S* knows-for-purpose-ϕ that *P*, then *S* knows-for-purpose-ϕ that *S* knows-for-purpose-ϕ that *P*"—but not across fragments. So fragmentation allows for a qualified defense of iteration principles.[31]

Mathematical and logical knowledge

In *The Construction of Logical Space* (2013: ch. 4), Agustín Rayo proposes a theory of mathematical knowledge in which fragmentation plays a central role. Rayo's goal is to answer the following question: "[H]ow should one model cognitive accomplishment in mathematics?" (Rayo 2013: 99). The core of Rayo's proposal is that

> cognitive accomplishment in logic and mathematics can be modeled, in part, as the acquisition of *information transfer abilities*: abilities whereby information that was available for the purposes of one set of tasks becomes available for the purposes of a different set of tasks. (Rayo 2013: 102)

The notion of fragmentation is then applied to spell out what information transfer is:

> A subject who has access to a piece of information for some purposes but not others is usefully thought of as being in a *fragmented* cognitive state. … One can model an information-transfer ability as the instantiation of a relation of *accessibility* amongst different fragments within the subject's cognitive state. (Rayo 2013: 104)

Rayo offers several examples, one of which is the following. Consider a farmer *A* who has to buy tiles to cover the area of her piece of land and also buy fencing for its perimeter. The farmer might know that she has to buy enough tiles to cover 81 square meters (the area of her land) and that (say) the patch is square-shaped while being unable to calculate how much fencing to buy.

[30] For Greco, "fragmentation" strategies actually encompass both what we've called horizontal and what we've called vertical fragmentation (see Section 3.1). So another way to respond to the unwitting historian is to ascribe two different kinds of knowledge attitudes to the unwitting historian, one for her first-order knowledge of English history and another for her second-order knowledge (cf. Greco 2015: n. 9 and p. 9).

[31] In addition, Greco (2014) argues that fragmentation also solves a puzzle about cases of "epistemic akrasia," in which a subject has a belief of the form "*P* but I oughtn't believe that *P*."

On Rayo's model, the farmer's lack of mathematical knowledge and deductive skills does not amount to lack of information. The farmer does possess the information that the area is 81 square meters, which entails that the perimeter is 36 meters. However, the information she possesses is not accessible for the task of buying fencing; it is only accessible for the task of buying tiles. Once the farmer acquires the required knowledge and skills, thus making the calculation, the information she already had becomes accessible for the new task (as well as further tasks, such as answering the questions "What is the perimeter of the patch?"). No "new" information has been acquired, but "old" information has become more accessible. Mathematical accomplishments, then, are often achieved by increasing the accessibility of information that is "fragmented" and unavailable for the task at hand. Stalnaker's early appeal to fragmentation is thus implemented and developed within a theory of mathematical knowledge (Rayo himself recognizes that his account draws on some of Stalnaker's ideas; Rayo 2013: 97).[32]

Questions and beliefs

Yalcin (2018) also goes back to Stalnaker's discussion of fragmentation. Yalcin notes that the so-called "problem of logical omniscience" (see Sections 2.1 and 4.2 above) is actually a set of related problems, and fragmentation seems to help with some but not all of them. In particular, even a fragmented possible worlds model will still exhibit the following problematic property:

> **Closure under entailment (fragmented)**
> If *A* bears the belief relation to *p* with respect to some state, and *p* entails *q*, *A* bears the belief relation to *q* with respect to some state. (Yalcin 2018: 29)

To see why this is problematic, consider one of Stalnaker's (1984) examples. William III of England believed that England could avoid war with France. If England could avoid war with France, then England could avoid nuclear war with France. However, it seems that William III did *not* believe that England could avoid nuclear war with France. The problem is that, since the first proposition entails the second, all worlds where England avoids a war are worlds where England avoids a nuclear war. So the fragmented possible worlds model entails that, relative to some belief state (or "fragment"), William III believed that England could avoid nuclear war with France. This is not what we want, for William III simply had no such belief—there should be *no* belief state/fragment relative to which he believed the proposition in question.

While taking fragmentation on board, Yalcin also develops the possible worlds model further to address this problem. The main idea is to think of belief as "question-sensitive." Consider a

[32] This brief summary only touches on a few aspects of Rayo's theory—Rayo also provides a formal model of fragmented cognitive states, discusses belief reports, and offers independent motivation for his view.

question as a “partition of logical space” (Yalcin 2018: 32) that divides the space of worlds into cells. Each cell is a class of worlds that “yield the same answer” to the question, although they might differ in other respects. So each cell corresponds to a possible complete answer to the question. The idea, then, is that states of belief are relativized to questions—or, as Yalcin also puts it, to “resolutions of logical space” (Yalcin 2018: 31–3).

Now consider our problem case again. “Will there be a war?” is one question; “Will there be a nuclear war?” is a different question.[33] In particular, in some worlds there will be no nuclear war, but there will be war nonetheless (naval war, say). Therefore, the two partitions/questions are not identical. William’s belief state is defined relative to the first question—the question partitions logical space into two cells, and William’s belief state excludes exactly the worlds in one of them, the one where there will be war.[34] All the worlds where there will be a nuclear war are thereby also excluded. But now consider the nuclear war question, which also partitions logical space into two cells. In addition to excluding the worlds in the nuclear war cell, William’s belief state also excludes several others—naval war worlds, for instance. William’s belief state is thus too rough and unspecific with respect to nuclear war matters; it is therefore undefined relative to the nuclear war question. This explains why William III did not believe that a nuclear war could be avoided, although he did believe that a war could be avoided. This brief sketch does not do justice to all the complex details of Yalcin’s view, but it should give an idea of this recent interesting application of the fragmentation hypothesis.

5. Open questions and overview of the volume

In this final section, we provide an overview of the contributions to the volume while also presenting some open questions that constitute promising avenues for future research and discussion about fragmentation. Each contribution is related to a group of questions, to locate it more clearly within the debate and to allow the reader to navigate more easily between the volume’s contents.

Fragmentation: Foundational issues and motivation

One crucial question for anyone who defends fragmentation is whether there is enough reason to do so. As previously highlighted, the rival model of unity enjoys widespread support and has important virtues. Is a departure from the model justified?

This question takes center stage in Daniel Greco’s contribution to the volume, “Fragmentation and Coarse-Grained Content.” As Greco shows—and as we explained in Section

[33] We changed the two questions slightly with respect to Stalnaker’s example so as to simplify the discussion.

[34] Of course, William will also have several other beliefs which rule out further worlds. But we can set this complication aside here.

4.2—fragmentation can be used within possible worlds models of content to address the problem of logical omniscience (what Greco calls “the problems of coarse grain”). This leads to a natural worry: Is fragmentation an ad hoc maneuver designed to save a specific theory? Greco thinks not: Fragmentation is independently motivated in virtue of “certain very attractive views about the relationship between beliefs, desires, and action” and as a “‘modest’ approach to model-building in philosophy” (Greco, Chapter * in this volume: *).

In their “Fragmentation and Information Access,” Adam Elga and Agustín Rayo also aim to provide support for fragmentation, while at the same time developing a specific version of the general approach. Their main thesis is that “in order to account for the behavioral dispositions of a subject with imperfect access to her information, we need to specify what information is available to her relative to various purposes” (Elga and Rayo, Chapter * in this volume: *). The authors offer a model to represent information accessibility for a subject and apply that model to puzzling phenomena like confusion and imperfect recall.

In their “The Fragmentation of Belief,” Joseph Bendana and Eric Mandelbaum defend and develop the fragmentation idea through a thorough examination of empirical evidence from cognitive science. While recognizing that a unified model of belief storage is popular and *prima facie* attractive, the authors criticize the model as conflicting with recent findings on a variety of cognitive phenomena. They then propose a set of empirically-informed hypotheses that better fit the data and serve as the core for a fragmented model of belief storage.

Andy Egan’s “Fragmented Models of Belief” is a self-proclaimed “advertisement for the program of constructing fragmented models of subjects’ propositional attitudes” (Chapter * in this volume: *). Egan lays out the reasons that motivate abandoning unified models of attitudes in favor of fragmented ones and provides a roadmap of questions and issues for future research on such fragmented models.

As we go on to explain in what follows, questions about the motivation for fragmentation are by no means abandoned in the rest of the volume—other chapters also identify novel and interesting sources of support, suggesting that one of fragmentation’s main virtues might well be its versatility in dealing with a diverse range of philosophical puzzles.

Rationality and fragmentation

Another set of open questions concerns the normative status of being in a fragmented belief state. One way to inquire about the rationality of being fragmented is this: “Is it always rationally better, cognitive capacities permitting, to be unified than fragmented?” (Egan, Chapter * in this volume: *). If we accept *Consistency* and *Closure* as global requirements on an agent’s overall belief system, the answer seems to be yes: To be fragmented, it is often assumed, *is* to fall short of perfect rationality (Stalnaker 1984). At the other end of the spectrum, Egan answers no: “[F]ragmentation can serve as a useful damage-control device in cases where agents have belief-forming mechanisms that are liable to go wrong. The fact that agents can have such fallible belief-forming mechanisms can make it more rational, in certain kinds of cases, to be fragmented than to be unified" (Egan, Chapter * in this volume: *; cf. Egan 2008). If it's sometimes but not always more rational to be fragmented than unified, the question about the rationality of fragmentation shifts: When, or in what ways, is it rational to be fragmented? According to Cherniak (1986), an agent with limited cognitive resources should compartmentalize her memory in a way that makes the search of her stored beliefs, or memory, most efficient. As we have seen (Section 4.1), this will occasionally lead the agent to make mistakes, but these do not impair her status as a "minimally rational" subject.

In her "Rationality in Fragmented Belief Systems," Cristina Borgoni provides another answer to questions about the rationality of fragmentation. She argues that while consistency remains a rationality requirement within fragments, it's not a rationality requirement across fragments. Instead, Borgoni proposes responsiveness to evidence as an inter-fragment rationality requirement for distinguishing between irrational and rational cases of fragmentation. Seth Yalcin, in his "Fragmented but Rational," also denies that coherence is an inter-fragment rationality requirement on an agent's overall belief state and rhetorically asks: "[W]hy isn't this [applying a coherence requirement across fragments] just confusedly projecting a constraint that applies to individual fragments onto the entire doxastic state?" (Chapter * in this volume: *). Going beyond coherence, Yalcin searches for other inter-fragment requirements of rationality that are violated by "fragmentation per se." His search coming up short, he concludes that "fragmentation per se is not a failure of rationality" (*).

Fragmentation and language

Section III takes up issues concerning the relation between fragmentation and language. The idea of fragmentation is generally stated as a hypothesis about the structure and organization of our mental states. But can this hypothesis also have interesting consequences and applications at the level of language? This question is explored in the two chapters in this section.

In his "Fragmentation and Singular Propositions," Robert Stalnaker presents two puzzles concerning necessary a posteriori truths and propositional attitudes, respectively. His proposal is to address the puzzles by combining the strategies of fragmentation and "diagonalization." These strategies were originally proposed in Stalnaker's *Inquiry* (1984); in his contribution to this volume, Stalnaker sheds new light on the relationship between fragmentation and diagonalization, explaining why they are not competing but complementary strategies and pointing toward applications to further philosophical puzzles.

Dirk Kindermann's "On the Availability of Presuppositions in Conversation" proposes applying fragmentation to another problem arising in connection with linguistic communication. Having introduced the notions of presupposition and common ground as they are generally construed in the pragmatic tradition of Stalnaker and Lewis, Kindermann argues that these notions need refinement. What speakers presuppose—what they mutually take for granted for the purposes of conversation—is available or unavailable to them only relative to a conversational task at hand. Kindermann argues that traditional conceptions of presupposition are unable to explain this phenomenon, and he proposes a different approach, where the common ground is modeled as a fragmented information state.

Fragmentation and mental files

The chapters in Section IV examine the relationship between fragmentation and the so-called "mental files" hypothesis.[35] While the idea of mental files has received a lot of attention in the recent literature, there has not been a thorough examination of the relationship between fragmentation and mental files. Are these competing hypotheses? Are they complementary? Are they different ways of implementing the same basic idea? These questions are taken up in the two chapters in this section.

In his "Do Mental Files Obey Strawson's Constraint?" François Recanati focuses on "Frege Cases"—cases where a subject is unaware or mistaken about some identity fact. Recanati argues that, within his version of the mental files framework, Frege Cases should be treated as an instance of fragmentation. Recanati's argument is based on his defense of "Strawson's constraint"—an influential and controversial principle about how to model our knowledge of identities. Once this basic principle is in place, Recanati argues, we can look at mental files as a specific manifestation of our fragmented cognitive architecture.

Michael Murez's "Belief Fragments and Mental Files" begins by presenting a cluster of philosophical issues which, he argues, should be addressed by combining fragmentation and mental files in a single account. Murez then presents a dilemma for both fragments and files: Both hypotheses "threaten to be either explanatorily lightweight or empirically refuted" (Murez, Chapter * in this volume: *). He contends that we should embrace the second horn of the dilemma, opening fragmentation and mental files up to empirical refutation or confirmation.

Murez's dilemma is in fact part of a broader set of questions having to do with the descriptive adequacy of fragmentation theories. Should we interpret these theories as idealized or normative models, or are they meant to provide a correct description of our actual cognitive processes? If the latter is true, does this mean that fragmentation theories should be subject to the possibility of empirical confirmation or refutation, as Murez argues?

There has been some discussion of these questions in the previous literature. For instance, Norby (2014) examines recent work on memory in cognitive psychology; focusing on "cue-dependent retrieval," Norby says "it suggests that fragmentationalism may not find motivation

[35] There is a large body of literature on the topic—see Recanati (2012, 2017) for an influential recent statement of the mental files model.

from the perspective of psychological science" (Norby 2014: 33). More generally, Norby's claim is that

> although empirical psychology provides a number of ways in which we might divide beliefs up by psychological type, it is unlikely that any of them are suitable for fragmentationalism's purposes. (Norby 2014: 32)

Questions about the empirical adequacy of fragmentation receive in-depth discussion in this volume (see the chapter by Bendana and Mandelbaum and the chapter by Murez), thus providing new inputs to future discussion about this important issue.

Fragmentation and implicit attitudes

In Section V, the discussion focuses on implicit attitudes and their connection with the fragmentation hypothesis. Implicit attitudes have received a great deal of discussion in the recent philosophical literature. This is not surprising, given that they are not only an important object of study for philosophy of mind and cognitive science but also a phenomenon of great social and political significance. As noted in Section 2.6, psychological research has brought to light striking cases of implicitly biased attitudes towards groups such as "African-Americans, women, and the LGBTQ community" (Brownstein 2019). Gaining a better understanding of implicit attitudes thus turns out to be essential on both theoretical and practical grounds. In this section of the volume, the issue is tackled from a number of different perspectives, with special attention being paid to the extensive empirical literature in this area.

Can fragmentation help to shed light on the nature of implicit bias, and implicit attitudes more generally? In Section 2.6, we briefly sketched one way to apply fragmentation to the problem, focusing on Brownstein's (2019) example of implicit bias towards women. Note, however, that there is at least one important difference between implicit bias and some classic examples of fragmentation: Implicit bias can be remarkably insensitive to evidence and can escape cognitive control on the subject's part. A person can be aware of their own bias against a particular group, believe that their prejudiced attitude is incorrect and harmful, and desire to change the attitude in question, yet the implicit attitude will often persist.[36] Other cases of fragmentation are prima facie different—in describing his own case, for instance, Lewis tells us that his beliefs changed "straightway" once he noticed the inconsistency (see Section 2.2). When trying to apply fragmentation to this domain, one will therefore have to account for the specific features that seem to distinguish implicit bias from other cases of fragmentation.

[36] For an example, see Schwitzgebel's (2010) example of "Juliet the implicit racist."

One theorist who integrates fragmentation within his theory of implicit bias is Joseph Bendana, in his chapter “Implicit Attitudes Are (Probably) Beliefs.” Against views that distinguish implicit attitudes from beliefs, Bendana presents a dilemma: “The criteria of belief they invoke are either too strong to be criterial for explicit beliefs or too weak to exclude implicit attitudes from the category of belief” (Chapter * in this volume: *). On the other hand, there is a body of empirical data that is often cited against the view that implicit attitudes are beliefs; to address this challenge, Bendana appeals to an “independently motivated, fragmented model of the mind” (*), which can be combined with Bendana’s view of implicit attitudes to explain the relevant data.

Josefa Toribio’s contribution “Implicit Bias and the Fragmented Mind” also reviews a large body of empirical literature, but it criticizes the attempt to apply fragmentation to cases of implicit bias. In the first part of the paper, Toribio presents a dilemma for fragmentation views: Assuming a “dispositionalist” account of belief, fragmentation is explanatorily vacuous; assuming a “representationalist” account, fragmentation is irrelevant, with the notion of “access” playing the central explanatory role. In the second part of the paper, Toribio focuses on implicit bias and offers an alternative reading of the phenomenon, arguing that “a representational, contextualist, non-fragmentationalist, and affect-laden account of the dissonance between implicit and explicit biases” is superior to a fragmentation-based explanation.

In her chapter “Rational Agency and the Struggle to Believe What Your Reasons Dictate,” Brie Gertler focuses on one of the issues mentioned earlier: the sensitivity—or lack thereof—of our attitudes to evidence and cognitive control. The chapter argues against “agentialism,” an influential view that holds that rational agency has special normative significance. Gertler argues against one particular claim held by agentialists, namely that “[t]he normative significance of rational agency transcends its instrumental value as an especially effective means of achieving alignment between attitudes and reasons” (Chapter * in this volume: *). Gertler presents a case where the subject succeeds in bringing her beliefs in line with her reasons but does so through an indirect process of belief shaping. The case is identified as an instance of fragmentation, where the subject has recalcitrant beliefs that resist the force of reasons.

Eric Schwitzgebel’s contribution, entitled “The Pragmatic Metaphysics of Belief,” defends a “pragmatic” approach to the nature of belief against what he calls an “intellectualist” approach. On the latter view, intellectual endorsement of a proposition is sufficient to ascribe a belief in that proposition to the relevant subject. Not so on a pragmatic view, where belief is “behaviorally demanding”: “[T]o really, fully believe, you must also ‘walk the walk’” (Chapter * in this volume: *). The chapter is based on a case of cognitive dissonance, which raises questions about what the subject really believes: “Does Daniel believe that the working poor deserve at least as much respect as those of higher social status?" (*). Schwitzgebel's chapter brings social and political issues to the fore, connecting them to the debate about implicit bias in an especially vivid way.[37]

References

Aronson, E. (1997), 'Back to the Future: Retrospective Review of Leon Festinger's "A Theory of Cognitive Dissonance"', in *The American Journal of Psychology* 110/1: 127–37.

Bilgrami, A. (2006), *Self-Knowledge and Resentment* (Harvard University Press).

BonJour, L. (1988), *The Structure of Empirical Knowledge* (Harvard University Press).

Borg, E. (2007), 'If Mirror Neurons Are the Answer, What Was the Question?', in *Journal of Consciousness Studies* 15/8: 5–19.

Borgoni, C. (2018), 'Unendorsed Beliefs', in *Dialectica* 72/1: 49–68.

Bratman, M. (1987), *Intention, Plans and Practical Reason* (Harvard University Press).

Brownstein, M. (2019), 'Implicit Bias', in *Stanford Encyclopedia of Philosophy* (Fall 2019 edition). URL: https://plato.stanford.edu/archives/fall2019/entries/implicit-bias/.

Cherniak, C. (1986), *Minimal Rationality* (MIT Press).

Christensen, D. (2004), *Putting Logic in Its Place: Formal Constraints on Rational Belief* (Oxford University Press).

Davidson, D. (1963), 'Actions, Reasons, and Causes', in *Journal of Philosophy* 60/23: 685–700.

—— (1973), 'Radical Interpretation', in *Dialectica* 27/3–4: 313–28.

—— (1982/2004), 'Paradoxes of Irrationality', in *Problems of Rationality* (Clarendon Press), 169–88.

—— (1986/2004), 'Deception and Division', in *Problems of Rationality* (Clarendon Press), 199–212.

—— (2001), *Essays on Actions and Events* (Oxford University Press).

—— (2004), *Problems of Rationality* (Clarendon Press).

Davies, M., and Egan, A. (2013), 'Delusion: Cognitive Approaches – Bayesian Inference and Compartmentalization', in K.W.M. Fulford et al. (eds.), *The Oxford Handbook of Philosophy and Psychiatry* (Oxford University Press), 689–727.

Dennett, D. (1981), 'True Believers: The Intentional Strategy and Why It Works', in A. Heath (ed.), *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford* (Clarendon Press), 150–67.

Easwaran, K., and Fitelson, B. (2015), 'Accuracy, Coherence, and Evidence', in T. Gendler et al. (eds.), *Oxford Studies in Epistemology 5* (Oxford University Press), 61–96.

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

[37] Thanks to Joseph Bendana and Cristina Borgoni for their very helpful comments, and special thanks to Cristina Borgoni for her extra work on the final section.

Fine, K. (2007), *Semantic Relationism* (Blackwell).

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

Frankish, K. (2010), 'Dual-Process and Dual-System Theories of Reasoning', in *Philosophy Compass* 5/10: 914–26.

Gendler, T. S. (2008a), 'Alief and Belief', in *Journal of Philosophy* 105: 634–63.

—— (2008b), 'Alief in Action (and Reaction)', in *Mind & Language* 23/5: 552–85.

Gertler, B. (2011), 'Self-Knowledge and the Transparency of Belief', in A. Hatzimoysis (ed.), *Self-Knowledge* (Oxford University Press), 125–45.

Greco, D. (2014), 'A Puzzle About Epistemic Akrasia', in *Philosophical Studies* 167/2: 201–19.

—— (2015), 'Iteration and Fragmentation', in *Philosophy and Phenomenological Research* 91/3: 656–73.

—— (2019), 'Fragmentation and Higher-Order Evidence', in M. Skipper and A. Steglich-Petersen (eds.), *Higher-Order Evidence: New Essays* (Oxford University Press), 84–104.

Griffiths, T., Tenenbaum, J., and Kemp, C. (2012), 'Bayesian Inference', in K. Holyoak and R. Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning* (Oxford University Press), 22–35.

Harman, G. (1986), *Change in View: Principles of Reasoning* (MIT Press).

Kaplan, M. (1996), *Decision Theory as Philosophy* (Cambridge University Press).

Kolodny, N. (2007), 'How Does Coherence Matter?', in *Proceedings of the Aristotelian Society* 107/1: 229–63.

—— (2008), 'Why Be Disposed to Be Coherent?', in *Ethics* 118/3: 437–63.

Lehrer, K. (1974), *Knowledge* (Oxford University Press).

Lewis, D. (1982), 'Logic for Equivocators', in *Noûs* 16/3: 431–41.

Mandelbaum, E. (2015), "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias', in *Noûs* 50/3: 629–58.

Norby, A. (2014), 'Against Fragmentation', in *Thought: A Journal of Philosophy* 3/1: 30–38.

Parikh, R. (1987), 'Knowledge and the Problem of Logical Omniscience', in *Proceedings of the Second International Symposium on Methodologies for Intelligent Systems* (North-Holland Publishing Co.), 432–39.

—— (1995), 'Logical Omniscience', in L*ogic and Computational Complexity. International Workshop LCC '94. Selected Papers* (Springer-Verlag), 22–9.

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

Recanati, F. (2012), *Mental Files* (Oxford University Press).

—— (2017), *Mental Files in Flux* (Oxford University Press).

Schwitzgebel, E. (2010), **'**Acting Contrary to Our Professed Beliefs, or The Gulf Between Occurrent Judgment and Dispositional Belief', in *Pacific Philosophical Quarterly* 91/4: 531–53.

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

—— (1991), 'The Problem of Logical Omniscience, I', in *Synthèse* 89/3: 425–40.

—— (1999), ‘The Problem of Logical Omniscience, II’, in *Context and Content* (Oxford University Press), 255–73.

van Fraassen, B. (1995), ‘Fine-Grained Opinion, Probability, and the Logic of Full Belief’, in *Journal of Philosophical Logic* 24: 349–77.

Yalcin, S. (2008), ‘Modality and Inquiry,’ PhD thesis, MIT.

—— (2018), ‘Belief as Question-Sensitive’, in *Phenomenology and Philosophical Research* 97/1: 23–47.