THESE CONFABULATIONS ARE GUARANTEED TO IMPROVE YOUR MARRIAGE!
Toward a teleological theory of confabulation

[a]Samuel Murray & [b]Peter Finocchiaro

[a]Psychology & Neuroscience Department, Duke University, Durham, NC, USA
[b]School of Philosophy, Wuhan University, Wuhan, China

Forthcoming in Synthese

1

Abstract

Confabulation is typically understood to be dysfunctional. But this understanding neglects the phenomenon's potential benefits. In fact, we think that the benefits of non-clinical confabulation provide a better foundation for a general account of confabulation. In this paper, we start from these benefits to develop a social teleological account of confabulation. Central to our account is the idea that confabulation manifests a kind of willful ignorance. By understanding confabulation in this way, we can provide principled explanations for the difference between clinical and non-clinical cases of confabulation and the extent to which confabulation is rational.

Keywords: Confabulation; Rationality; Clinical; Social Agency; Cognitive Bias; Willful Ignorance

Introduction

They say that not telling the truth is bad, especially in relationships. They say that trust is the foundation of a healthy romance. But what they say might only be half true. Healthy marriages and long-term relationships persist, in part, through confidentially endorsed idealizations. According to some studies, individuals greatly exaggerate the positive qualities of their romantic partners (Penton-Voak et al., 2007). Even starry-eyed optimists can concede that these people are not telling the whole truth. And yet couples self-report stronger feelings of being in love when they idealize their partner to some degree, even over the course of a long-term relationship (Miller et al. (2006)).

When trust and honesty seem so important, why do people resort to stating falsehoods about their partners? And why do these false claims seem to work? They work because such idealizations are rational confabulations. Or so we will argue.

Historically, confabulation was thought to be a clinical phenomenon. Some individuals would sincerely assert absurd claims that directly conflicted with the obvious. A right-hemisphere stroke patient, for instance, would report that their left index finger was pointed at the doctor's nose even though the patient's hand lay motionless (Ramachandran, 1996). But individuals without any apparent psychological or neurobiological impairment also seem to confabulate. One study had individuals compare the quality of identical pairs of nylon stocking pantyhose (Nisbett & Wilson, 1977; Wilson & Nisbett, 1978). Most participants rated the right-most pair as most preferable. When asked to defend this selection, many insisted they based their decision on that pair's superior quality, even though the only noticeable difference between the pairs was their relative position.

Psychologists, psychiatrists, and philosophers try to specify what is common to these and other cases of confabulation along several dimensions. Some focus on

2

the causes of confabulation. Some focus on similarities in the content of confabulations. And some focus on the physiological correlates of confabulation.

Many of these proposals presume that instances of *clinical* confabulation are paradigmatic of the phenomenon in general. The understanding of confabulation that emerges from this presumption is one where clinical cases command most attention. Because these cases are generally dysfunctional, the phenomenon of confabulation itself is seen as paradigmatically dysfunctional.

The psychological models that result from this understanding, however, do not reflect a unified conception of confabulation. For one, there are diverse neurobiological bases and physiological correlates of confabulation that crosscut its different manifestations (Bortolotti & Cox, 2009; Metcalf et al., 2007) and there appears to be no adequate content-based classification of confabulation (Coltheart, 2017). Moreover, dysfunctional accounts of confabulation have difficulty explaining the potential benefits of clinical confabulation (Bortolotti & Sullivan-Bissett, 2018).

Some more recent work on confabulation shifts the focus to non-clinical confabulation and its benefits. Non-clinical confabulation seems somewhat common, and this calls into question the presumptive paradigmatic status of clinical confabulation. By taking non-clinical confabulation as paradigmatic, we can develop a better model of confabulation, one that explains the occurrence of everyday, non-clinical confabulation, but also explains the occurrence of confabulation in clinical settings.

That is just what we will do in this paper. We begin with this alternative presumption, that non-clinical confabulation is paradigmatic. We then sketch a theory of confabulation on that basis. In particular, we start with the idea that non-clinical confabulation results from properly functioning cognitive mechanisms that contribute to achieving goods inherent to socially embedded agency. In this regard, our theory is teleological: it characterizes confabulation in terms of the functional role it must fulfill in order to facilitate the achievement of these goods.

The inadequacy of existing proposals might lead some researchers to be pluralistic about confabulation. They might say that there are no generally informative conditions that jointly define the category and that what we call "confabulation" is, in reality, a disjunction of psychological phenomena connected only by superficial relations of similarity. We want to resist this temptation. Pluralism runs the risk of conflating confabulation with other kinds of irrational beliefs or delusions. Such conflations would make it more difficult to develop practically useful diagnostic conditions and remediation programs (see Bortolotti & Cox, 2009: 956). Beyond this, we believe there is a unified story to tell about confabulation. This unification is an important upshot of our teleological account of confabulation, though the proof will be in the pudding.

Here's the plan for what's to follow. We first introduce the epistemic approach to confabulation and consider whether paradigmatic instances of confabulation are

dysfunctional. Then, to support our contention that paradigmatic instances of confabulation are not dysfunctional, we consider a recent proposal on its benefits (Bortolotti, 2018). Next, we argue that confabulation provides at least two socially oriented benefits: predictability and relational resilience. We use these benefits to develop a framework according to which confabulation manifests a kind of willful ignorance. Then, we apply our framework to explain the rationality of confabulation, provide a principled means of distinguishing between clinical and non-clinical cases of confabulation, and make concrete predictions about when confabulation is likely to occur.

1. *Extant Accounts of Confabulation*

In this section, we discuss some extant accounts of confabulation. We accept many elements of these accounts. But we also caution against some of their assumptions as well as their needless reliance on questionable empirical evidence.

1.1 A Critique of Confabulation as a Dysfunction

Confabulation was first discussed scientifically in the late nineteenth century, when the Russian psychiatrist Sergei Korsakoff began to document some odd patterns in his patients' conditions. These patients suffered from persistent "illusions of memory". One patient seemed to be unaware that she was ill, and even said that she went for a walk earlier despite her obvious immobility. Another patient seemed to completely fabricate a story about meeting a famous doctor the day before (Korsakoff, 1890). Subsequent research continued the theme of confabulation as a false memory disorder (see Schnider, 2017: ch. 2). Discussions were confined to clinical contexts and focused on persistent and fantastical confabulations in brain-damaged patients (Kopelman, 1987; Moscovitch, 1989).

The overarching implication of this research program is that confabulation is dysfunctional. Confabulating patients state obviously false things when they should say true things. But they don't even seem aware of the fact that they are stating obviously false things. Furthermore, the fact that patients consistently produce false memory reports supports the idea that confabulation results from malfunction. In light of this, many recent definitions of confabulation focus on false memories, neurobiological damage, or both (see Turner & Coltheart, 2010 for overview).

These accounts are instances of the two-factor theory of confabulation. Typically, a two-factor theory posits that confabulation arises from the interactions between two neurobiological factors. The first factor is damage to a perceptual or mnemonic process that produces ill-grounded (and often false) outputs. The second factor is damage to an executive monitoring process that would, when functioning correctly, manipulate and correct the outputs of the damaged perceptual or mnemonic process. Because the two-factor theory claims that

confabulation is the result of the interactions between damaged cognitive processes, it characterizes confabulation as dysfunctional.

There are at least three serious issues with the presumption that confabulation is a clinical memory disorder that manifests as a persistent tendency to produce false memory reports.

First, the content of a confabulation need not be false (see Bernecker, 2017). Sometimes an individual's confabulation merely misreports when an actual event happened (Borsutzsky et al., 2008). When asked, one Korsakoff patient said they spent last New Year's Eve with their wife in a restaurant. They had in fact spent a New Year's Eve in this way, but it was before the patient's hospitalization. That being said, the extent to which a confabulation is true is beside the point. Imagine that you were a Capgras patient who claimed that your son had been replaced by an imposter. What you said would be a confabulation even if – somehow – the person before you really were an imposter! This is because you lack evidence of the replacement. Thus, whether some verbal report counts as confabulation depends on the *justification* of the report, not its *truth-value*.

Second, there is non-clinical confabulation. Consider, for example, Nisbett and Wilson's aforementioned research on evaluations of consumer goods. Researchers asked participants to compare the quality of four identical pairs of nylon stocking pantyhose. A disproportionate number of participants chose the right-most pair. When asked, many participants denied that the relative position of the pairs of pantyhose played any role in which pair they chose. They insisted that the right-most pair appeared to be better. These individuals' claims are importantly similar to those made by individuals in "paradigmatic" clinical cases of confabulation. It is unlikely, however, that underlying neurobiological damage explains confabulation in consumer choice studies.

The third issue is that confabulation is not restricted to memory reports. In consumer choice cases, individuals supply *reasons* or *justifications* for their selections. These reasons probably do not match the reasons that actually motivated their behavior. But it would be misleading to say that these individuals are reporting a false memory of which reasons motivated their behavior. Rather, what they are doing is asserting that certain reasons (like superior quality) were motivating when those reasons probably were not motivating. This kind of confabulation is likely not restricted to artificially constructed studies of consumer behavior. Consider studies of bias in hiring decisions. When asked to evaluate identical resumes, individuals exhibit a preference for the resumes of people with white-sounding names over those with black-sounding names (King et al., 2006; Sullivan-Bissett, 2015). If they were asked to explain their decisions, they would likely claim that the differences were made on the basis of important, albeit small, differences in the candidates' qualifications. But such an explanation is likely incorrect. Absent any real differences between candidates beyond race, it seems likely that race-based considerations motivated the decision rather than differences in qualification.

Reasons-based confabulation can also be found in clinical cases. Consider, for instance, visual anosognosia patients who deny their blindness (Celesia & Brigell, 2005). When these patients are asked to prove that they are not blind (or, at least, to explain why they fail to perform simple visual discrimination tasks), they will sometimes cite poor environmental conditions or a lack of interest in performing the task (Kaufman & Milstein, 2013). Confabulations that pertain to an individual's reasons or justifications are rarely a matter of producing a false memory report.[1]

In response to these and similar criticisms, some researchers have developed theories that focus on the epistemic features of confabulation. One influential epistemic account defines confabulation partly in terms of ill-grounded thought. According to William Hirstein, some individual, S, confabulates with respect to some claim, p, if and only if:

1. S claims that *p*;
2. S believes that *p*;
3. S's thought that *p* is ill-grounded;
4. S does not know that their thought that *p* is ill-grounded;
5. S should know that their thought that *p* is ill-grounded;
6. S is confident that *p*.[2]

Hirstein's account entails that confabulation is dysfunctional insofar as it states that there is something the confabulating individual should do that they systematically fail to do.[3] This straightforwardly explains why clinical cases of confabulation often manifest as false memory reports despite not consisting entirely of false memory reports.

Hirstein's account does not entail that confabulation manifests only in clinical contexts.[4] It can also include both memory reports and justification reports as types of confabulation. According to his account, however, someone who

---

[1] Doris (2015: 141) also distinguishes between false memory reports and faulty justification reports. However, he calls these latter reports "rationalizations rather than confabulations, to distinguish them from the clinical phenomena" (2015: 141). We accept the distinction, but reject the implication that 'confabulation' refers exclusively to some clinical phenomenon.

[2] Hirstein, 2005: 5.

[3] To press this point, Hirstein also says: "Confabulation involves absence of doubt about something one should doubt…It is a sort of *pathological* certainty about ill-grounded thoughts and utterances" (2005: 4). Moreover, his account is rooted in the claim that confabulations results from frontal processes failing to operate properly (2005: 178). He takes this to be evidence that confabulation is a product of *systematic* failure. Finally, his sense of 'should' at issue is functional: "The use of 'should' here involves no more than the claim that an optimally functioning car should move down the road" (2005: 192).

[4] Hirstein supports his account, however, with appeals to elements that make confabulation seem inherently clinical. Confabulation is persistent (or pathological), contra-functional, and grounded in neurological damage.

confabulates fails to know something that they should know. So, confabulation is dysfunctional by definition. Someone who confabulates fails to live up to certain epistemic norms.

We take issue with this idea that confabulation is inherently (or paradigmatically) dysfunctional. We take issue with it because some research indicates that confabulations appear to be beneficial in both clinical (Bortolotti & Sullivan-Bissett, 2018) and non-clinical (Doris, 2015: ch. 6) contexts. This research calls into question the claim that confabulation is the result of an improperly or poorly functioning psychology.

In fact, we believe that confabulation is globally functional. We admit that the sorts of clinical cases that served as the historical origin of our understanding of confabulation are ones where the behavior is clearly dysfunctional. But it is a historical accident that such cases became the presumed paradigmatic instances of confabulation.[5]

The appearance of inherent dysfunction arises from focusing on such instances of confabulation in overly narrow settings. To properly understand the global functionality of confabulation, we must see how confabulation fits into the wider tapestry of human agency. In particular, we must see how confabulation operates in non-clinical contexts. As the relationship benefits mentioned at the outset suggest, there are many cases where confabulation contributes to positively functioning agency.

In line with what we have said here, some recent discussions have shed the presumption that confabulation is inherently dysfunctional by more carefully exploring its potential benefits. Coltheart (2017) discusses the possibility that confabulation results from the proper functioning of several cognitive processes. Sullivan-Bissett (2015: 555-57) argues that confabulation facilitates self-understanding by explaining behavior in terms of justifiable patterns of reasons. Bortolotti (2018) has recently highlighted additional practical and epistemic benefits of confabulation. (In the next section, we summarize the main parts of her account before suggesting some extensions.) On these views, confabulation is not inherently dysfunctional.

These accounts take the benefits of confabulation as central to constructing a theory of confabulation. We accept the central idea behind these alternatives, but we take issue with the specific claims they make about confabulation's benefits. In order to better situate our own account (in Sections 2 and 3), we will now evaluate these purported benefits.

1.2 Bortolotti and the Agency-Enhancing Benefits of Confabulation

---

[5]

   As we explain below, confabulation typically does most of its work behind the scenes. The easiest exceptions to find lie in certain clinical settings where confabulation blatantly reveals itself. This might explain why the study of confabulation initially focused almost exclusively on clinical cases of it. We thank a reviewer for raising this point.

Bortolotti's proposal regarding the benefits of confabulation rests on two claims: (i) certain factors threaten people's sense of self-integrity, and (ii) these threats are compensated for by confabulation's ability to enhance agency and preserve a sense of self-integrity. Thus, confabulation has epistemic and practical benefits (Bortolotti, 2018: 243).

Bortolotti discusses two potential threats to agency. The first comes from unconscious influences that seemingly drive a large amount of people's thoughts and actions. The second threat comes from the lack of direct access people have to their unconscious behavioral influences. According to Bortolotti, these two threats cause people to systematically misidentify the motivating reasons for their actions (Bortolotti, 2018: 242). They also create conflicts for people's intuitive sense of being competent, rational, and reasons-responsive (Bortolotti, 2018: 241). As a result, these factors undermine self-integrity and agency.

On Bortolotti's view, these potential threats to agency are all mitigated by confabulation. In general, confabulation enhances agency by justifying behavior in ways that conform to an integrated self-image. Bortolotti (2018) identifies three specific benefits of confabulation: (1) the construction of a better self, (2) the integration of self-related information, and (3) the maintenance of the social self. Some of these benefits might have downstream motivational effects. One way for someone to maintain the social self is to conceive of themselves as a reasons-responsive agent. Maintaining this self-conception might actually improve future reasons-responsiveness. Some benefits might also generate further epistemic benefits. By presenting themselves in a socially admirable light, a confabulating individual might improve their standing in a knowledge community and thereby make it easier to exchange information within that community.

Bortolotti acknowledges that not confabulating – i.e., achieving maximal accuracy in the identification and communication of justification for thoughts and actions – might lead to a form of self-integration. If people did not confabulate, they might form an integrated conception of self that accurately represents their decision-making as someone who is subject to numerous unconscious influences. However, a more accurate self-image might not provide the motivational and epistemic benefits afforded by confabulation. As Bortolotti claims:

> A better-grounded or even accurate explanation [instead of confabulating] is unlikely to play self-enhancing and self-integrating roles to the same extent as the rival [confabulating] explanation. Acknowledging that the consumer choice was not based on the quality of the items but on an unconscious tendency to favour items on the right-hand side may not support people's sense that they are competent agents and decision makers, and may not help them identify patterns that contribute to their construction of a coherent image of themselves as discerning consumers (Bortolotti, 2018: 244).

8

In short, then, the potential benefits of confabulation stem from the role confabulation plays in enhancing an individual's agency in the face of numerous threats to an integrated sense of themselves as a rational actor.

We think that Bortolotti's proposal makes important contributions to our understanding of confabulation's role in supporting and enhancing agency. In particular, we believe there is an account of socially embedded, temporally extended agency that explains why confabulation has the benefits that Bortolotti discusses.

While we aim to build on Bortolotti's proposal, we also take issue with some of the empirical support that Bortolotti cites. According to Bortolotti, confabulation generates the appearance of consistency among the kinds of things that might serve as motivating responses. It also aligns these motivating responses with an individual's underlying sense of who they are and what they care about. This is important because on her view, the preferences that are causally relevant to decision-making varies considerably in virtue of unconscious or situational influences:

> Preferences may vary considerably depending on the circumstances, and this is also due to such preferences being influenced by mental processes that cannot be controlled via deliberation or accessed via introspection, such as priming effects, basic emotional reactions, and implicit biases (2018: 242).

So, according to Bortolotti, confabulation is important because people's actual preferences are subject to a range of unconscious or situational forces (i.e. the priming effects, affective responses, and implicit biases mentioned above) and the influence of these forces cannot be "controlled via deliberation or accessed via introspection" (2018: 242).

The evidence for these claims has, however, recently been disputed. While there are many studies that purportedly find evidence of the influence of unconscious factors on behavior, meta-analyses of these studies have revealed only small or domain-specific effects on overt behavior (Demaree-Cotton, 2016; Landy & Goodwin, 2015; Oswald et al., 2013). Thus, we lack clear empirical evidence about the magnitude of influence that unconscious factors have on the deliberative components of behavior.

Beyond this failure to obtain clearly domain-general results, the research itself has been subject to conceptual challenges. Some researchers note that the studies used to support claims about unconscious influences on behavior have boundary conditions that greatly limit generalizability (see Mele, 2009; Miller, 2013; May, 2018; Newell & Shanks, 2014). This reinforces our skepticism about the credentials of the empirical research.

The extent to which unconscious factors influence people's behavior remains an open question. For that reason, we are hesitant to rely on it in developing a theory of confabulation. This empirical caution does not *impugn* Bortolotti's

account. Our point is that this issue is largely *irrelevant* to the broader issue of confabulation's role in a properly functioning cognitive life. Confabulation would be beneficial even if it turned out that people's motivations were not subject to widespread unconscious influences inaccessible to their introspection. Bortolotti's main point about confabulation's positive status can make do with more modest empirical claims. To see that, however, we need to first understand the social context of confabulation as well as the underlying picture of agency that gives rise to it.

2. *The Social Teleology of Confabulation*
In the next two subsections, we discuss two circumstances under which people tend to confabulate. What is common to these two circumstances is that their confabulatory behavior is essentially social and (arguably) rational. They thus serve as illustrative examples of the properly functional nature of confabulation. We consider the benefits of confabulating in these circumstances, like boosting predictability and relational resilience.

2.1 The Confluence of Random Selection and Predictability

Let's return to consumer choice.[6] Researchers have found that people confabulate reasons for their purchases. In consumer choice studies, individuals are asked to choose from among a set of options with no readily discernible difference between them. When asked to justify their selection, participants confabulate reasons based on quality or appearance. Why do people do this?

We suggest that consumer choice cases demonstrate one species of confabulation, one where the confabulation is the product of an interaction between two processes. Each process plays a crucial role in our psychic economy that is largely independent of the other's. But there are specific circumstances where the two collide. When they do, they produce confabulations.

Define a *Buridan's Ass scenario* as a scenario in which someone must decide between options with no readily discernible difference. For most people, such scenarios are a routine part of life. For instance: they routinely choose which pump to use at the gas station, or which box of pasta to pick up at the grocery store. These scenarios do not involve decisions between gas station chains or brands of pasta. They involve decisions between particular pumps at the same station and particular boxes of De Cecco (or Barilla, or whatever brand of pasta you might happen to favor).

6

   Studies that investigate positional effects on choice also tend to elicit confabulation (Bar-Hillel, 2015). Hall et al. (2012) identified similar effects for political and moral opinions. Participants first completed surveys on their political and moral opinions. Researchers then reversed some answers and ask participants to explain their (now reversed) opinions. Nearly 70% of participants defended the reversed position, even when the level of initial agreement tended toward the poles (i.e. strong agreement or strong disagreement). These studies and consumer choice studies are importantly similar: in them, people justify choices by appealing to reasons that do not discriminate the chosen option from other available options.

In these scenarios, people must make a choice – they can't just leave empty handed! But it would require a lot of effort to fully evaluate every option. For example, it would take several minutes to measure individual boxes of pasta to discern a minute weight difference that might justify selecting one box over the others. The value provided by that slight weight difference is exceeded by the comparatively large value of the time spent weighing. More generally, in Buridan's Ass scenarios the payoff from making the "optimal" selection is almost never worth the effort required to discern the relevant reasons. It would therefore be highly inefficient for someone to try to discern reasons that would rationalize their decision.

In these scenarios, people typically don't care about making the "optimal" selection. But because they have to make a choice, people often intentionally routinize the choices they make in these types of scenarios. Some might, for example, always grab the closest box of pasta, starting from the shelf closest to eye level and moving from left to right, like reading a book. But people are likely to encounter new and unexpected Buridan's Ass scenarios. In the absence of an established routine, it is beneficial – for the sake of efficiency – to have a random-selection mechanism that tips the balance in favor of one option.[7] Thankfully, people seem to have this (see Icard, Forthcoming). This mechanism is not consciously cultivated, and in fact its operations are largely invisible to the individual. Nevertheless, the benefits to efficiency that the mechanism provides establish its rationality in Buridan's Ass scenarios.[8]

This random-selection mechanism is the first process that underlies the species of confabulation exhibited most clearly in consumer choice cases studies.

The second process relates to the adaptive value of predictability. For organisms that engage in repeated social interactions, predictability yields a variety of benefits (McNamara et al., 2008). One illustrative benefit of predictability is the promotion of specialization within social niches, a crucial factor in joint activity and the development of communities (Bergmüller & Taborsky, 2007). This specialization enables people to occupy particular roles that reinforce expectations about how they will behave across a variety of circumstances. Predictability is thus an important component of our presentation as individuals who are capable of productive and fruitful social engagements. Hence, there is adaptive value in having mechanisms that produce seemingly predictable and apparently non-random patterns of behavior.

---

[7]
  In this context, random selection need not be the activity of a mechanism, strictly speaking. Random selection might be the upshot of processes with certain computational properties that can be modeled as continuous random diffusion processes (Ratcliff & McKoon, 2008). For simplicity, though, we heuristically talk in terms of mechanisms.

[8]
  In our view, the mechanism is arbitrary in that it responds to some reasons and not to others in a manner that is unprincipled. This mechanism thus does not exacerbate Bortolotti's worry about unconscious influence undermining reasons-responsiveness. Though the selection mechanism exhibits a kind of arbitrariness, this arbitrariness is compatible with reasons-responsiveness.

When combined, these two processes produce something unexpected. Think of some mundane item from the grocery store that you bought this week, just like you do nearly every week. Now imagine being asked why you got that *specific* item. The real answer to their question might involve some complicated and arbitrary computational process that underlies the operations of your reasons-responsive mechanisms.[9] But we suspect that wouldn't be the answer you'd instinctively give (it certainly wouldn't be the answer we'd give!). That kind of arbitrariness works against people's desire to appear predictable in social environments. Hence, we find that Buridan's Ass scenarios generate a conflict between the outputs of the socially beneficial mechanisms that produce the appearance of non-randomness and the outputs of optimally engineered mechanisms that mediate random choices. This conflict sometimes produces a confabulation.

Our claim is not that confabulation *guarantees* the distinctive benefit of apparent predictability in Buridan's Ass scenarios. Our claim, rather, is that these two distinct psychological processes sometimes conflict. People in these scenarios *attempt* to secure the benefits of both apparent predictability and randomized choice without facing the seemingly inevitable consequences of the conflict between the two processes. In their attempt to cover their behavior in a socially acceptable way, they produce confabulations.

2.2 Marriage Illusions as Confabulations

People in long-term, committed relationships try to support their decisions to maintain their relationships with marriage illusions: projected rationalizations that fit socially acceptable patterns of justification. These marriage illusions positively affect the individual's perceptions of and attitudes toward their partner in a long-term committed relationship (Conley et al., 2009). We think that these illusions demonstrate another species of confabulation.

Some studies have found that people claim that their partner is more attractive relative to the assessments of others (Penton-Voak et al., 2007; Swami et al., 2007). One study found that married people tend to be unaware of any faults in their partner (Edmonds, 1967).[10] Another study found that, with respect to certain interpersonal traits, partners were consistently rated by their significant others more positively than they rated themselves in self-assessment (Murray et al., 1996). And a 13-year longitudinal study found that married people tend to find their partner more agreeable than expected (Miller et al., 2006). These results indicate that individuals tend to present their choice of partner as more obvious

---

9

More fully: reasons-responsive mechanisms are composed of computational processes that, in certain situations (such as these Buridan's Ass scenarios), bias decision-making toward one option from a number of practically indiscriminable options. This is still the activity of a reasons-responsive mechanism because the output is a choice made on the basis of reasons (even though those reasons might not be discriminating).

10

This illusion also holds for LGBTQ and cohabiting heterosexual couples (Conley et al., 2009).

than it (in all likelihood) actually was. We acknowledge that there are important differences between long-term relationships and consumer choices. People don't normally pick from an array of simultaneously available and indistinguishable partners, after all. But the selection of a romantic partner is, to a non-negligible degree, arbitrary. Rather than admit to this arbitrariness, people attempt to justify it in socially fitting ways (see Doris, 2015: 141).

Marriage illusions and consumer choice rationalizations are similar insofar as they both are attempts to justify decisions for which the individual lacks discriminatory justification. But they are also importantly different. For one, consumer choice cases involve one-off short-term selections, whereas long-term relationships require an initial choice followed by persistent commitment. We therefore suggest that there are further reasons for confabulation in relationships that are not found in consumer choice.

We think that marriage illusions facilitate persistence by boosting resilience. One study reports that positive illusions are nearly universal among individuals satisfied with their marriage or long-term relationship (Fowers et al., 2001). One such illusion concerns the probability of divorce. The likelihood of divorce shifts between 40% and 60% (subject to variation relative to different demographic factors; see Wilcox and Marquardt, 2011). However, when asked to self-assess, individuals consistently underrate the likelihood of their own relationship ending in divorce (the mean estimation is 22.6%, while the modal estimate is 0%; see Fowers et al., 2001). Any honest (or realistic) person should see that these assessments are probably ill-grounded. The fact that people nevertheless consistently make inaccurate self-assessments strongly suggests that their reports are confabulations of a sort.

Both kinds of idealizations (i.e. those concerning the decision and those concerning the commitment) manifest a form of ignorance. But, again, they serve a social purpose. Marriage confabulations bolster relational resilience. Couples report being more in love when they idealize their partner (Miller et al., 2006). Relationships that indulge in a degree of positive illusion tend to be healthier and more satisfying in the long-run. Perhaps this is because these positive illusions function like performance demands that make people behave more positively toward their partner (Doris, 2015: 133). An individual's idealizations of their partner affect their attitudes toward that partner. These attitudes in turn raise the likelihood of the long-term success of the relationship. The relational optimism behind these confabulations might protect or bolster the long-term stability of a relationship, particularly in difficult times. As John Doris aptly remarks: "…inflated optimism may help sustain a union. In the dark hours of domesticity, could one muster the gumption to tough it out while vividly aware that it's only a coin flip the bloody toil will pay off?" (Doris, 2015: 132). Accurate, even near-accurate, assessments would likely fail to provide these benefits.[11]

---

[11]
   This phenomenon extends beyond cases of long-term, committed relationships. Many endeavors benefit from increased resilience resulting from a (potentially overinflated) sense of

2.3 Agency and Confabulation

In Section 1.2, we suggested that Bortolotti's claim about the beneficial role of confabulation could get by on more empirically modest credentials. We are now able to develop that suggestion more fully. In doing so, we situate our account of confabulation in a broader theory of socially embedded agency. This, in turn, explains the functional profile of confabulation and helps to unify the two species of confabulation we have discussed thus far.

How people view themselves and present themselves to others plays an important role in their agency. At a minimum, their self-conception mediates the scope of deliberative options, the availability of certain reasons, and the relevant norms of rationality. The roles someone occupies, the aspirations they have, and the narrative details that bring these together partially underwrite the exercise of agency (see Velleman, 2009).

What an individual's self-conception fully amounts to is an interesting question, but one that is beyond the scope of this paper. Roughly speaking, though, we think that an individual's self-conception has two components. One component is descriptive, while the other is aspirational. The descriptive component represents the individual's view of who they currently are. The aspirational component represents what a person aims to become and what projects they aim to realize. An individual's self-concept is likely composed of many different kinds of psychological attitudes. But for the purposes of this paper we are interested in the beliefs constitutive of the descriptive component (see Murray & Vargas, 2020 for further discussion of the role of self-concept in agency).

While people's self-conception plays a useful role in facilitating self-understanding and guiding behavior, it also plays a useful role in coordinating with others. When someone projects themselves as a certain kind of agent (e.g., as a teacher or as an athlete) they help others to predict how they will behave cross-situationally, and thereby help to form expectations about how they will interact with others (Chen, Chen, and Shaw, 2004; Swann & Read, 1981).

When people help others to understand their behavior, they facilitate important forms of collaboration. Collaboration is a central feature of socially embedded agency. While we cannot offer a full theory here, we sketch this collaborative view of agency as a way of explaining the benefits we ascribe to confabulation (for fuller defenses of this view, see Doris, 2015; Tomasello, 2016; Velleman, 2009).

Human beings are saddled with strict limitations. Some of these derive from anatomy; others are psychological in kind. The precise source of these limits is unimportant. What is important, though, is that these strict limits stand in stark contrast to the lofty ambitions that many of us have. Many of life's most

---

competence. The reason for this becomes clear when situated within our account of socially embedded agency in the next section.

14

meaningful pursuits and projects are too complicated to carry out alone. In order for people to pursue the things they care most about, then, they must collaborate with others as a way to compensate for their strict limitations.

Collaboration is a unique kind of cooperative endeavor. Collaboration requires joint intentionality, where each collaborator can see the other collaborators as participants in the joint activity. Even further, collaborators must see how the others see themselves as participants in the joint activity. They would otherwise struggle to set others up for success (Tomasello et al., 2012). Collaboration is a cognitively rich enterprise that requires people to predict what their collaborators need and how they will respond. This complexity explains the need for predictability that we mentioned earlier. Collaboration works best when everyone acts in highly predictable ways (van Dellen et al., 2015).

Collaborative ventures are typically complex tasks that unfold over time. Early in our evolutionary history, for instance, hunters had to collaborate for days on end in order to bring down a woolly mammoth (Tomasello et al., 2012). While modern humans undertake different projects, they face similar problems. The joint activity often does not serve the local (short-term) interests of all participants. This problem is especially prevalent when the joint activity is undertaken in the service of some abstract goal or ideal whose reward structures are non-obvious. For the collaboration to succeed, there needs to be a way to manage commitment over time. This need, we claim, is met through resilience.

More specifically, resilience works on two fronts. For an individual, resilience helps to martial their own motivational resources in service of maintaining commitment over time. For a group of collaborators, resilience helps to increase the perceived commitment among them, and thereby helps to maintain a willingness to continue collaborating.

We have argued that these collaborative ventures require predictability and resilience. We suggest that these goods are, in part, secured by confabulation. But why confabulate to secure these goods, rather than through some other, seemingly less dysfunctional, process? Because confabulation produces (or maintains) useful forms of *ignorance*. In Buridan's Ass scenarios, an individual's confabulation produces (or maintains) an ignorance of the arbitrariness behind how choices are made in situations with practically indiscriminable options. In marriage illusions, an individual's confabulation produces (or maintains) an ignorance of the accurate (and sometimes grim) prospects for long-term relationship success. This ignorance is useful because other features of the individual's agency might otherwise undermine predictability and resilience in ways that threaten their ability to collaborate.[12]

---

[12]
   On this view, confabulation is an essentially social act because it is communicative (where the communication produces or maintains ignorance). This communicative element of confabulation distinguishes it from delusion, since delusion can occur without communication. We thank a reviewer for indicating this consequence of our view.

We think the empirical support for this role of confabulation is more secure than the empirical support for the roles mentioned by Bortolotti. Unconscious influences on people's decisions may undermine their sense of agency, though the evidence on this matter is unclear.[13] The question is far from settled., In contrast, collaboration is undoubtedly a central element of social agency. In outlining the role of confabulation in collaboration, we provide better empirical support for the claim that confabulation is a product of a properly functioning psychology. We do not deny that there is empirical support for Bortolotti's claim that there is psychological value to having a coherent and largely positive conception of the self. In fact, we are inclined to think that there is. However, this claim does not differentiate Bortolotti's account of confabulation from the one that we develop below because our account can also explain why such a conception would be psychologically valuable.

As we noted earlier, there are adaptive benefits to randomness. People are limited, and in particular people are limited in their capacity for self-control. Localized instances of randomness generate the appearance of *unpredictability.* In these situations, people preserve their overall, global, sense of predictability through confabulation. Such confabulation generates ignorance that can boost resilience and thereby counteract the negative effects of limited self-control, like the appearance of unpredictability.

There are also good reasons to engineer a highly predictable system with the capacity for long-term commitment. Without predictability and long-term commitment, it would be difficult to engage in different kinds of collaborative projects. That being said, there are good reasons to engineer an unpredictable system with limited self-control.[14] To get the best of both worlds, human beings need a system that can strategically produce ignorance about *many* of the facts that describe them. This need brings to light the precise sense in which confabulation is rational. Confabulation, considered globally, is expected to help satisfy more of our rational desires than alternative, non-confabulatory, behaviors. Confabulation is instrumentally rational because of the demands of socially embedded and temporally extended agency.[15]

13

There is some evidence that people take consciousness to be necessary for free agency, though this research does not distinguish between free agency requiring some form of consciousness and particular *exercises* of free agency requiring conscious mental activity in its immediate proximal etiology (Shepherd, 2015). Other evidence suggests that people easily confuse unconscious motivational factors with mechanistic causes of action (where mechanism is incompatible with agency; see De Brigard, Mandelbaum, & Ripley, 2009). Thus, there appears to be no conclusive evidence for or against the claim that knowledge of the widespread influence of unconscious processes on decision-making would undermine individual sense of agency.

14

To note just two examples of the benefit of unpredictability: (1) it is beneficial to be unpredictable when engaging in predator avoidance (See Richardson et al., 2018) and (2) it is beneficial to be unpredictable when engaging in creative thinking (Beaty et al., 2016).

15

Our claim about the rationality of confabulation is similar to Bratman's external sense of rationality in intentional action (Bratman, 1987: 43). From the external view, intentional action

We have used consumer choice cases and marriage cases to illuminate the functional role of confabulation. We have focused on consumer choice cases in particular because of the role these cases have played in the empirical study of non-clinical confabulation. But are these cases generalizable? If they are not, then our theory of confabulation has limited explanatory value.

Issues about generalizability are difficult to settle from the armchair. People certainly seem to encounter such cases from time to time. But (1) whether they encounter them frequently or infrequently and (2) whether they encounter them in a broad range of contexts or only in highly specific ones are empirical issues we cannot settle at this moment. However, we think that our explanation of consumer choice cases provides a model for understanding the role of confabulation in how people navigate Buridan's Ass scenarios.

Similarly, our explanation of marriage cases might seem of limited utility. After all, how many endeavors share the motivational perils of marriage?

On reflection, though, marriage cases might be more generalizable than Buridan's Ass scenarios. Many long-term commitments, especially those that are greatly valued, are such that we are tempted to abandon them from time to time. This raises a problem of how we can form rational commitments in light of the real possibility of temptation and back-sliding (see Marušić, 2015). In short, we think that confabulation is an important mechanism for shielding individuals from the evidence that might undermine the possibility of rational long-term commitment in epistemically respectable ways. The full story of how this works is undoubtedly more complex. However, this at least suggests that many long-term commitments (and not just marriages) rely on confabulation to preserve the rationality of persisting in them over time. Thus, marriage cases provide a useful model for understanding how confabulation works in various temporally extended endeavors. To name just a few intriguing comparisons: resolving to quit smoking, committing to a financial savings plan, dieting, or learning to play a musical instrument.

By understanding confabulation as a process that contributes to socially embedded agency, we can gain a better sense of the benefits that confabulation aims to confer. This understanding also reveals the central role of *producing ignorance* that defines the functional profile of confabulation. In the next section, we elaborate on this idea further and more precisely state what confabulation is.

3. *Confabulation as a Manifestation of Willful Ignorance*

---

is rational when and only when such action contributes to the satisfaction of rational desires. The rationality of desires is a more complicated issue that goes beyond the scope of this paper. We note briefly our preference for accounts of rational desire that state explain the rationality of desire in terms of how well grounded the desire is in appropriate experiences of the desirability characteristics of some state of affairs that forms part of the object of the desire (Audi, 1985).

Thus far, we have discussed a few of the circumstances under which people will non-dysfunctionally confabulate as well as the benefits that they stand to gain by doing so. In this section, we provide a more fully developed and more general account of confabulation. Confabulation, we argue, is best understood as a manifestation of willful ignorance. By understanding confabulation in this way, we can accomplish five theoretically significant tasks. First, we can make sense of the role that confabulation plays in collaborative agency. Second, we can explain the varieties of non-clinical confabulations. Third, we can articulate the differences between clinical and non-clinical confabulation while preserving its unity as a psychological phenomenon. Fourth, we can justify the ways in which confabulation can be rational. Finally, we can make testable predictions regarding the conditions under which we can expect people to confabulate.

3.1 The Willful Ignorance Framework

Let's begin by considering some curious empirical results. Surprisingly, people can be pushed to back off from their marriage confabulations. In one study, Solomon and Vazire (2014) found that individuals in romantic relationships inflate the attractiveness of their partner relative to the assessments of others, but that they also accurately report their partner's *reputation* for attractiveness (cf. Doris, 2015: 133). It seems, then, that marriage confabulators are not completely blind to the justificatory status of their confabulations. Solomon and Vazire speculate that people internalize both their own assessments and the assessments of others. When prompted, people can bring the latter to mind, and thereby tacitly acknowledge that their own assessments are inflated. Plausibly, then, confabulations like those found in marriage illusions result from the suppression of known, accessible information.

If that is the right read of the situation, then it appears that confabulation manifests a kind of willful ignorance.[16] Suppose that S states that *p*, where this statement is a confabulation. S's stating that *p* manifests willful ignorance when:
   (1) The statement is not based on a suitably large number of members in the information set {I} possessed by S, and
   (2) It is the case that if S were to access a suitably large number of members in {I}, S would not state that *p*, and
   (3) S fails to access members in {I} because the information is detrimental to realizing S's commitments, goal, and cares.

Condition (1) captures the fact that confabulating manifests ignorance. Condition (2) captures the fact that the ignorance is motivationally efficacious. And Condition (3) captures the fact that the ignorance is willful. We turn now to a

---

16

   Sometimes, willful ignorance is the result of a deliberate failure to acquire evidence (for instance, the suspicious spouse might be willfully ignorant of her partner's infidelity because she deliberately avoids checking her spouse's phone for evidence). The kind of willful ignorance manifested in confabulation is not like this. As we'll discuss below, the kind of willful ignorance manifested in confabulation is the result of a deliberate failure to access information that is either known or easily inferred given what is already known.

more explicit discussion of what the will is and how it relates to the willful ignorance manifested in confabulation.

An individual's will is a hierarchically-organized capacity structured by what the individual cares about. These cares establish volitional constraints that define an individual's limits of practical deliberation, including what they take to be potential reasons for action (Frankfurt, 1989). The volitional constraints in turn make certain options or commitments *unthinkable* (see Frankfurt, 1988).

The structure of an individual's volitional hierarchy sensitizes them to certain considerations and desensitizes them to others. As the structure becomes more entrenched (or central to the individual's identity) this sensitization might result in blindness to certain sets of considerations. For example, an individual's deeply entrenched care and commitment toward their family might set a volitional constraint such that they cannot even deliberate about the possibility of betraying their siblings for a trivial sum of money. Additionally, in the context of a long-term relationship, care and commitment toward their partner might make infidelity unthinkable. These structures provide an architecture for practical reasoning and deliberation that frames certain options as open while leaving other options out entirely.

The kind of willful ignorance manifested in confabulation is the result of a deliberate failure to access accessible information that is either known or easily inferred given what is already known.[17] The failure, however, is rooted in the individual's volitional structure insofar as the failure is in service of maintaining a commitment, goal, concern, or other volitional element. This willful element is crucial; without it, we would struggle to distinguish a confabulation from an unwittingly false statement.

As we understand it, deliberate unawareness is not intentional. We do not think that someone prone to marriage confabulations first has in mind a body of accurate relationship information and then forms an intention to ignore it. Rather, that individual's cares and commitments shape volitional constraints that form the background conditions for their subsequent confabulations.[18]

17

 This formulation might seem paradoxical. How can information that is practically unthinkable be accessible? While the information is unthinkable relative to the current context, this does not rule out the possibility of the information being accessed in a different context. Evidence that this kind of information is *available* comes from Solomon & Vazire's (2014) study, where people were able to back off of their confabulations in certain contexts.

18

 While our account borrows heavily from Frankfurt's conception of the will, it does not depend on his being the correct account *of the will*. We think that these structures (cares, goals, commitments, volitional constraints, etc.) are typical of human psychology. The relationship between these structures and confabulation is independent of whether the concept of the will can be fully analyzed in terms of these structures. This brings out the close relationship between our account and other motivational accounts of confabulation (Sullivan-Bissett, 2015; Coltheart, 2017). Our account adds to these an underlying notion of agency that supplies a framework for thinking about the functional role of non-clinical confabulation.

Understanding these volitional constraints is crucial for understanding deliberate unawareness. An individual's will is a capacity structured by the hierarchy of their cares, concerns, and commitments. These cares establish volitional constraints that define the individual's limits of practical deliberation, including what they take to be potential reasons for action (Frankfurt, 1989). These volitional constraints make certain options or commitments *unthinkable* (see Frankfurt, 1988).

Now that we've discussed the general connection between the will and confabulation, we will briefly discuss how it manifests in our paradigmatic cases of properly functional confabulation.

People typically care about their relationships, especially their long-term, committed relationships. These relationship-directed cares set volitional constraints that make certain considerations typically unthinkable – for example, that their partner is not attractive. In some contexts, the set of available considerations may include only considerations the awareness of which would fundamentally disrupt the long-term health of the relationship. As a result, people might exclude information about their partner that is, in reality, both relevant and accurately representative.

Importantly, though, this is not a merely incidental fact about marriage illusions. While the benefits gained (e.g., relational resilience) might be peculiar to marriage illusion confabulation, the same psychological framework can be used to explain other species of properly functional confabulation. Consider confabulation in Buridan's Ass scenarios. As we explained in Section 2.3, people care about presenting themselves as predictable because it facilitates collaborative ventures. Consequently, people willfully ignore information about the randomness that underlies some of their more mundane choices. Again, we expect there to be some interesting differences across these circumstances in precisely how and why confabulation occurs. But they remain unified by their shared overall function – namely, as contributing toward maintaining the kind of willful ignorance necessary for the acquisition of practical goods inherent to socially embedded agency.

3.2 Applying the Willful Ignorance Framework

People can, in the right setting, back off from their confabulations. This fact might seem to suggest that confabulation is an intentional activity where people decide when and where they want to confabulate. But confabulation is not an intentional activity. If confabulation were intentional, it would be difficult to distinguish it from lying. Nevertheless, confabulation does have the appearance of intentional activity (cf. Fotopoulou, 2009). We think that any adequately developed theory of confabulation must explain this appearance while simultaneously preserving confabulation's status as a non-intentional activity.

The Willful Ignorance Framework provides just such an explanation. The framework characterizes confabulation as a manifestation of a kind of willful

ignorance. As we explained above, willful ignorance results from an individual's volitional constraints. People never intentionally act to set their volitional constraints. Volitional constraints set the boundaries for deliberation. Because they are the grounds of choice making, they are not the targets of our choices. An individual can, of course, deliberate about their cares or commitments. But the end result of such a deliberation is not the formation of an intention. When an individual revises their final ends, they are doing something other than revising their intentions (see Richardson, 1994 for further discussion). While the resulting unawareness is not intentional, because it is shaped by the individual's will (i.e. the current structure of their cares) – the unawareness is *deliberate*.

So why does confabulation *appear* to be intentional? Solomon and Vazire's (2014) results show that willful ignorance is susceptible to *calibration*. In calibration, an individual experiences a shift in their volitional constraints that result from a reorganization of their volitional hierarchy that is itself based on a change in the available reasons or the balance of the individual's values. There are some considerations that people are normally blind to because those considerations negatively bear on the relationships that they care about.[19] But their care for the relationship can be temporarily eclipsed by other cares. For example, when a researcher prods an individual to assess the reputation of their partner (especially in private, anonymous contexts), that individual's care for productive social engagement can momentarily eclipse their care for the relationship. As a result, that individual accesses information that accurately represents their partner. The fact that willful ignorance can be calibrated explains why an individual can be induced to make accurate reports rather than confabulate, even though this change in behavior is not the result of anything intentional.[20]

19

 Some examples suffice to bring this out. We've touched on how this kind of blindness affects long-term romantic relationships. But such blindness might also affect particular religious communities that include members who have done terrible things (e.g. the sexual abuse scandal in the Catholic Church). Commitment to the community might generate blindness to these occurrences as a way to maintain the commitment. Similarly, sports fans, in order to continue enjoying their sport, might blind themselves to unfair labor practices or player safety concerns. Fans of American football, for example, might blind themselves to the dangers of head trauma and CTE that have recently been linked to the sport.

20

Two additional pieces of evidence support the claim that confabulation can be calibrated (and, hence, that the difference between clinical and non-clinical confabulation is the possibility of calibration). First, spontaneous clinical confabulation is associated with executive dysfunction, particularly with damage to the ventromedial frontal cortex and the orbitofrontal cortex (Gilboa et al., 2006; Turner et al., 2008). These regions also subserve various cognitive control processes like salience processing (Boorman, Rushworth, & Behrens, 2013). The associations between spontaneous confabulation and cognitive control (and the lack of cognitive control) salience provide one reason to believe that confabulating requires a failure of control that manifests a form of ignorance. Second, some mechanistic models of confabulation posit that confabulation partially results from a failure to monitor and suppress thoughts with false or ill-grounded contents (Johnson & Raye, 2000; Turner & Coltheart, 2010). These monitoring and inhibitory components are also constituents of self-control and cognitive control that ground capacities to act deliberately and intentionally (see Davisson & Hoyle, 2017). These failures of control, however, can be modulated by the agent's personal-level concerns, values, and plans (Kool, Shenhav, & Botvinick, 2017). This empirical evidence is useful for understanding the possibility,

Calibration's role in non-clinical confabulation further highlights the agential aspects of confabulation. Our understanding of confabulation as a complex manifestation of agency helps to explain the way that the multiple realizability of confabulation coheres with the available empirical evidence regarding the many putative causes of confabulation. Some researchers highlight how confabulation results from lack of affective guidance in source monitoring and retrieval of memories (Hirstein, 2005: 180). Without the correct emotional constraints, people cannot adequately respond to social information. They might, as a result, confabulate. Other researchers note that confabulation can result from people's desires to fill gaps in memory, their desires to construct coherent self-narratives, or simply their desires to project their own wishful thinking (Fotopoulou, 2009). The Willful Ignorance Framework readily incorporates these psychological factors. The structure of an individual's volitional hierarchy partly determines their motivations, and emotions help to regulate thought and action in light of this hierarchy.

There are several implications here worth mentioning. First, our understanding of confabulation suggests that there is no unique neurological basis of confabulation and, hence, no single (or informatively unitary) cause of confabulation.[21] The multiple realizability of confabulation coheres with the available evidence regarding the numerous putative causes of confabulation. Some researchers highlight how confabulation results from lack of affective guidance in source monitoring and retrieval of memories (Hirstein, 2005: 180). Without the correct emotional constraint, people cannot adequately respond to social information. They might, as a result, confabulate. Others note that confabulation can result from people's desires to fill gaps in memory, their desires to construct coherent self-narratives, or simply their desires to project their own wishful thinking (Fotopoulou, 2009). Our framework incorporates both psychological factors. The structure of an individual's volitional hierarchy partly determines their motivations, and emotions help to regulate thought and action in light of this hierarchy.

Second, our understanding of confabulation suggests that clinical confabulation manifests a breakdown of agency. Despite this breakdown of agency, we can still understand confabulation in terms of the teleological aims of paradigmatic non-clinical confabulation (we will return to this point later). The Willful Ignorance Framework makes this clear by supporting the following principled distinction. Non-clinical confabulation is a manifestation of a properly functioning psychology. Clinical confabulation, on the other hand, provides a window into an individual psychology that is in some way defective. Importantly, these defects need not be traceable to any particular psychological process or structure;

---

which we discuss more below, that the difference between clinical and non-clinical confabulation consists in calibration.

21

   The qualification is meant to acknowledge that there might be some abstract categorization that unifies all relevant neurobiological substrates of confabulation. However, the level of abstraction might be so high as to make the corresponding category uninformatively vague (see Klein, 2012).

clinical confabulation might result from an overly rigid volitional hierarchy or an insensitivity to changing social information, both of which are psychologically holistic features. This division also explains the fact that non-clinical confabulation is sometimes difficult to detect but clinical confabulation is typically obvious (Chrobak & Zaragova, 2009). In non-clinical settings people tend to produce confabulations that fit an acceptable pattern of reasons and justifications (Coltheart & Turner, 2009: 185-86; Wheatley, 2009: 209-10). But, in clinical settings, people tend to produce confabulations that fail to exhibit sensitivity to shifts in patterns of reasons, and therefore fail to justify their statements. Similarly, in non-clinical settings, people are typically able to withdraw previous confabulations or to withhold from confabulating altogether. Clinical confabulators are typically unable to do so. As a result, non-clinical confabulation often raises fewer red flags.

Someone might take this characterization to suggest that what separates clinical from non-clinical confabulation is the complete absence of the ability to calibrate. We think that's too strong. Evidence suggests that clinical confabulators can back off from their confabulations, albeit only with great difficulty and very briefly (cf. Gassell & Williams, 1963; Mattioli et al., 1999). Hence, we propose that confabulation is a continuum. Toward one end of the continuum are confabulations that are agency-enhancing, including paradigmatic non-clinical confabulations like marriage illusions. Toward the other end are confabulations that reveal a defective sense of agency and a complete inability to calibrate, including paradigmatic clinical confabulations like those produced by Capgras patients. What marks a confabulation's position along the continuum is the extent to which it reveals the ability to calibrate. At one extreme is a perfectly fluid capacity for calibration, while the other extreme reflects a total deficiency in calibrating.

The Willful Ignorance Framework is compatible with other accounts of confabulation. Consider, for example, mnemonic theories of confabulation that focus on false memory reports and the mechanisms that underlie these reports (Burgess & Shallice, 1996; Schnider et al., 2000). Such theories are compatible with our framework because the mechanisms they implicate can be interpreted as manifesting different kinds of willful ignorance. Causal theories of confabulation define mnemonic confabulation in terms of an individual's seeming to remember that p failing to be causally related to some past representation of its being the case that p (see Bernecker, 2017; Robins, 2016). On the Willful Ignorance Framework, the past representation of it being the case that p might be one of (if not the only) item in the relevant informational set that the individual fails to access in confabulating.[22] The Willful Ignorance Framework is also compatible with process-based account of confabulation (see Robins 2020;

22

    Sometimes, a confabulating individual might report its being the case that p where it was never the case that p. In this case, it's not simply that the past representation does not play the right causal role in producing the state of remembering; rather, there is no such thing that could play this sort of causal role. However, it might still be the case that the individual is confabulating, because its not being the case that p might be easily inferred from other information accessible to the individual.

Michaelin, 2016). Being willfully ignorant of some information is a process that is causally relevant to producing confabulation. This marks a point of agreement between the Willful Ignorance Framework and accounts of confabulation that focus on mnemonic confabulation.

Finally, because confabulation partly results from willful ignorance, an individual confabulation can be more or less rational in virtue of the aptness of the confabulation relative to the volitional hierarchy of the confabulator and the available reasons. We thus offer the following schema for rational assessment: an agent produces a confabulation that is rational relative to their volitional constraints if and only if the confabulation supports responding to or ignoring considerations in ways that conform to the volitional hierarchy at the time the agent produces the confabulation. In this way, the rationality of confabulation is congruent with reasons-based accounts of justification and rationality (cf. Scanlon, 1998: 22-25). With this formulation of the rationality of confabulation, we can explain why confabulation is subject to normative assessment even though it is not an intentional action. We can also explain the sense in which clinical confabulation usually seems to be irrational: such confabulations typically fail to cohere with the individual's volitional constraints.[23]When an amputee confabulates that the obviously lost limb is still there, it is hard to reconcile that confabulation with what the amputee cares most about. In contrast, when a husband confabulates about the unmatched desirability of his wife, that confabulation (typically, though not always) fits quite nicely with what the husband cares most about.

The Willful Ignorance Framework also makes some empirical predictions about the conditions under which people are likely to confabulate. As we suggested earlier, people are more likely to confabulate in contexts where they care about what they're doing. This reality makes confabulation tricky to test in experimental settings (especially since our view accounts for the desire to appear predictable to others). However, one experimental possibility is to manipulate whether statements are made publicly or anonymously. This manipulation should significantly modulate rates of confabulation (controlling for perceived importance of the statement topic). For example, one possibility would be to collect first- and third-party assessments on the likelihood of sticking to a training program from people who recently signed up to complete a marathon. Our account makes two predictions: (1) there should be a small but significant effect of anonymity, and (2) the effect of anonymity will be greater for those who rate completing the marathon as more important relative to those who rate it as less important. This is because care positively correlates with the likelihood of confabulating. Anonymity mitigates the urge to confabulate, so anonymity should reduce the frequency of confabulation. Thus, individuals who don't care

---

[23] Clinical confabulations can help protect meaningful components of an individual's self-conception (Gunn & Bortolotti, 2018) and compensate for the individual's compromised reputational capital (Bortolotti, 2018). Nevertheless, confabulation that approaches paradigmatic clinical confabulation likely fails to achieve the social benefits of paradigmatic non-clinical confabulation.

much about the marathon should respond more or less the same with respect to predicting commitment, whereas people who do care should respond differently under conditions of anonymity. This indicates how experimenters might apply this account to the empirical study of confabulation outside of clinical settings (and outside the context of long-term committed relationships).

3.3 A Summary of Our Account

Let's first consider what confabulation *isn't* on our account. Confabulation is not intentional deception (see Hirstein, 2005: 28; Sullivan-Bissett, 2015: 553). In fact, confabulation isn't intentional at all. Confabulation is not a deficiency of human psychology. And while confabulation sometimes takes the form of a false memory report, this by no means exhausts the various forms that it can take.

Confabulation arises in virtue of people's volitional constraints. Confabulation manifests a kind of willful ignorance characterized by its unique explanatory and motivational aims. But confabulation is not inherently dysfunctional; paradigmatic non-clinical confabulation manifests a complex exercise of social agency. For example, confabulation often promotes the goods of predictability and resilience by generating a distinctive form of ignorance in the confabulator and their "audience".

There can be dysfunctional forms of confabulation. Some of these may constitute clinical forms of confabulation. We believe that clinical confabulation occurs when the individual has difficulty exercising their ability to calibrate their behavior in light of their volitional hierarchy. The ensuing confabulations can indeed be worrying and worth treating. But they are instances of the same kind of psychological phenomenon that is found in more mundane confabulations, like those examined in consumer choice studies. It is only because of the historical focus on clinical confabulation that such cases were taken as paradigmatic of the phenomenon in general. According to the framework that we have developed, the order of explanation should go in the other direction. Dysfunctional clinical cases should be understood as cases where an otherwise properly functioning confabulatory psychology goes awry.

One final challenge for the Willful Ignorance Framework is that it appears to make all non-clinical confabulations instances of self-deception.[24] But we think that the two can be distinguished in the following way. When someone confabulates with respect to p they must be confident that p. It is possible, however, for someone to be self-deceived with respect to p and not be confident that p. Confidence is thus one distinguishing feature between self-deception and confabulation (cf. Hirstein, 2005: 214). Furthermore, self-deception is sometimes intentional, whereas confabulation is never intentional (cf. Hirstein, 2005: 214-15).

24

   Admittedly, some might not see this as a problem. Hirstein (2005: 226) thinks that an individual self-deceived about whether p (and who confidently and wholeheartedly believes that p) who asserts that p is confabulating.

So, while we admit that confabulation and self-deception share key features, they are nonetheless distinct constructs.

Recall Hirstein's epistemic theory, which states that S confabulates that p if and only if:

1. S claims that *p*;
2. S believes that *p*;
3. S's thought that *p* is ill-grounded;
4. S does not know that their thought that *p* is ill-grounded;
5. S should know that their thought that *p* is ill-grounded;
6. S is confident that *p*.

We've argued that confabulation is a manifestation of a kind of willful ignorance. This idea conflicts with two aspects of Hirstein's account, namely conditions (4) and (5). Confabulators, in certain circumstances, seem to know (or seem to be in a position to know) that their confabulations are ill-grounded. So, condition (4) is at best misleading. Furthermore, condition (5), when applied to cases of non-clinical confabulation, is either trivial or false – depending on how we understand 'should'. If we understand 'should' in the all-things-considered sense, then condition (5) is false. As we argued above, there are practical reasons that justify confabulating on particular occasions, and the balance of those reasons might provide decisive reason for an agent to confabulate. If we understand 'should' in the epistemic sense, then condition (5) is trivial. There are a number of things that agents like us should do from the epistemic perspective. But we regularly fail to live up to these epistemic demands for practical reasons. For example, we should, epistemically, be Bayesians with respect to belief revision and updating. But our cognitive limitations and the demands for efficiency tend to outweigh these epistemic considerations. Perhaps people should, epistemically, know that their thoughts are ill-grounded on some occasions of confabulating. But, given the practical reasons in favor of confabulating, we might fail to live up to these epistemic demands or find them outweighed by other consideration.

In light of these considerations, we offer the following two-factor epistemic definition. S confabulates that p if and only if:

1. S claims that *p*;
2. S believes that *p*;
3. S's thought that *p* is ill-grounded;
4. S willfully ignores the fact that the thought that *p* is ill-grounded;
5. S is confident that *p*.

This definition counts as a two-factor theory because two factors are primarily responsible for confabulation: the process that produces the ill-grounded thought and the processes that are implicated in willful ignorance. The thoughts that are produced are ill-grounded because they fail to have appropriate epistemic justification. But, so long as confabulation exhibits appropriate reasons-

responsiveness, the ill-groundedness of the expressed thought does not preclude the confabulation from being rational.

Our definition blends elements from Hirstein, Coltheart, Sullivan-Bissett, and Bortolotti's views on confabulation. It borrows from Hirstein the skeleton of an epistemic theory. But it adapts it to better incorporate Bortolotti, Sullivan-Bissett, and Coltheart's idea that the core of confabulation consists in its teleological aims.

Conclusion

In this paper, we presented a novel account of confabulation that focuses on the role confabulation plays in the lives of social creatures living under constraints of bounded rationality. While some instances of confabulation seem strange or dysfunctional, on the whole confabulation is an adaptive process that results from a conflict between things that matter to the individual confabulating. Confabulation, on this view, manifests a kind of willful ignorance. Confabulation is thus connected to other reasons-based accounts of rationality. In this way, our account makes sense of the fact that confabulation can sometimes be rational or beneficial. We also offered a principled division between clinical and non-clinical cases of confabulation, drawn in terms of whether or not the individual's confabulating behavior is sensitive to both the individual's cares and the available practical reasons that justify confabulating.

References

Audi, R. 1985. "Rationality and Valuation," In G. Seebass and R. Tuomela (Eds.) *Social Action* (Dordrecht: Reidel), 243-77.

Bergmüller, R. and Taborsky, M. 2007. "Adaptive Behavioural Syndromes Due to Strategic Niche Specialization," *BMC Ecology* 7:12.

Bernecker, S. 2017. "A Causal Theory of Mnemonic Confabulation," *Frontiers in Psychology* 8:1207. doi: 10.3389/fpsyg.2017.01207

Boorman, E.D., Rushworth, M.F., and Behrens, T.E. 2013. "Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice," *Journal of Neuroscience* 33, 2242-2253.

Borsutzsky, S., Fujiwara, E., Brand, M., and Markowitsch, H.J. 2008. "Confabulations in Alcoholic Korsakoff Patients," *Neuropsychologia* 46, 3133-3143.

Bortolotti, L. 2018. "Stranger than Fiction: Costs and Benefits of Everyday Confabulation," *Review of Philosophy and Psychology* 9:2, 227-249.

Bortolotti, L. and Cox, R.E. 2009. "'Faultless' Ignorance: Strengths and Limits of Epistemic Definitions of Confabulation," *Consciousness and Cognition* 18: 952-965.

Bortolotti, L. and Sullivan-Bissett, E. 2018. "The epistemic innocence of clinical memory distortions," *Mind & Language* 33:3, 263-279.

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

Burgess, P.W. and Shallice, T. 1996. "Response suppression, initiation, and strategy use following frontal lobe lesions," *Neuropsychologia* 34:4, 263-72.

Celesia, G.G. and Brigell, M.G. 2005. "Cortical Blindness and Visual Anosognosia," *Handbook of Clinical Neurophysiology* 5, 429-440.

Chen, S., Chen, K. Y., and Shaw, L. 2004. "Self-verification motives at the collective level of self-definition," *Journal of Personality and Social Psychology* 86, 77-94.

Chrobak, Q.M. and Zaragova, M.S. 2009. "The cognitive consequences of forced fabrication: Evidence from studies of eyewitness suggestibility," In W. Hirstein (ed.) *Confabulation*.

Coltheart, M. 2017. "Confabulation and Conversation," *Cortex* 87, 62-68.

Coltheart, M. and Turner, M. 2009. "Confabulation and Delusion," In Hirstein (ed.) *Confabulation*.

Conley, T.D., Roesch, S.C., Peplau, L.A., and Gold, M.S. 2009. "A Test of Positive Illusions Versus Shared Reality Models of Relationship Satisfaction among Gay, Lesbian, and Heterosexual Couples," *Journal of Applied Social Psychology* 39:6, 1417-1431.

Dalla Barba, G., Cipolotti, L., and Denes, G. 1990. "Autobiographical memory loss and confabulation in Korsakoff's syndrome: a case report," *Cortex* 26:4, 525-34.

De Brigard, F., Mandelbaum, E., and Ripley, D. 2009. "Responsibility and the brain sciences," *Ethical Theory and Moral Practice* 12:5, 511-24.

Demaree-Cotton, J. 2016. "Do Framing Effects Make Moral Intuitions Unreliable?" *Philosophical Psychology* 29, 1-22.

Doris, J. 2015. *Talking to Our Selves* (Oxford: Oxford University Press).

Edmonds, V.H. 1967. "Marital Conventionalization: Definition and Measurement," *Journal of Marriage and Family* 29:4, 681-88.

Fotopoulou, A. 2009. "Disentangling the motivational theories of confabulation," In Hirstein (ed.) *Confabulation*.

Fotopoulou, A. 2010. "The Affective Neuropsychology of Confabulation and Delusion," *Cognitive Neuropsychiatry* 15:1-3, 38-63.

Fowers, B.J., Lyons, E., Montel, K.H., and Shaked, N. 2001. "Positive illusions about marriage among married and single individuals," *Journal of Family Psychology* 15:1, 95-109.

Frankfurt, H.G. 1987. "Identification and Wholeheartedness," In F.D. Schoeman (ed.) *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (New York: Cambridge University Press).

Frankfurt, H.G. 1988. "Rationality and the Unthinkable," In *The Importance of What We Care About* (New York: Cambridge University Press).

Frankfurt, H.G. 1989. "Concerning the Freedom and Limits of the Will," *Philosophical Tiopics* 17:1, 119-130.

Frankfurt, H.G. 1992. "The Faintest Passion," *Proceedings and Addresses of the American Philosophical Association* 66:3, 5-16.

Gilboa, A., Alain, C., Stuss, D.T., Melo, B., Miller, S., and Moscovitch, M. 2006. "Mechanisms of spontaneous confabulations: A strategic retrieval account," *Brain* 129, 1399-1414.

Gunn, R. and Bortolotti, L. 2018. "Can delusions play a protective role?" *Phenomenology and the Cognitive Sciences* 17, 813-33.

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.

Hirstein, W. 2005. *Brain Fiction* (Oxford: Oxford University Press).

Icard, T. Forthcoming. "Why Be Random?" *Mind*.

Johnson, M.K. and Raye, C.L. 2000. "Cognitive and brain mechanisms of false memories and beliefs," In D.L. Schacter and E. Scarry (eds.) *Memory, Brain, and Belief* (Cambridge, MA: Harvard University Press), 35-86.

Kaufman, D.M. and Milstein, M.J. 2013. "Visual Disturbances," *Kaufman's Clinical Neurology for Psychiatrists* (Elsevier), 261-85.

King, E.B., Madera, J.M., Hebl, M.R., and Knight, J.L. 2006. "What's in a Name? A Multiracial Investigation of the Role of Occupational Stereotype in Selection Decisions," *Journal of Applied Social Psychology* 36:5, 1145-59.

Klein, C. 2012. "Cognitive Ontology and Region- versus Network-Oriented Analyses," *Philosophy of Science* 79:5, 952-60.

Kool, W., Shenhav, A., and Botvinick, M.M. 2017. "Cognitive control as cost-benefit decision-making," In *The Wiley Handbook of Cognitive Control*, 167-89.

Kopelman, M.D. 1987. "Two Types of Confabulation," *Journal of Neurology, Neurosurgery, and Psychiatry* 50, 1482-1487.

Kopelman, M.D. 2010. "Varieties of Confabulation and Delusion," *Cognitive Neuropsychiatry* 15:1-3, 14-37.

Korsakoff, S.S. 1890. "Über eine besondere Form psychischer Storung, combinirt mit multipler Neuritis," *Arch Psychiatr* 21, 669-704.

Landy, J.F. and Goodwin, G.P. 205. "Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence," *Perspectives on Psychological Science* 10, 518-36.

Marušić, B. *Evidence & Agency: Norms of Belief for Promising and Resolving* (Oxford: Oxford University Press).

May, J. 2018. *Regard for reason in the moral mind* (Oxford: Oxford University Press).

McNamara, J.M., Barta, Z., Fromhage, L., and Houston, A.I. 2008. "The coevolution of choosiness and cooperation," *Nature* 451, 189-192.

Mele, A. 2009. *Effective Intentions* (Oxford: Oxford University Press).

Metcalf, K., Langdon, R., and Coltheart, M. 2007. "Models of Confabulation: A Critical Review and a New Framework," *Cognitive Neuropsychology* 24:1, 23-47.

Michaelin, K. 2016. "Confabulating, Misremembering, Relearning: The Simulation Theory of Memory and Unsuccessful Remembering," *Frontiers in Psychology* 7:1857. doi: 10.3389/fpsyg.2016.01857.

Miller, C. 2013. *Moral Character* (Oxford: Oxford University Press).

Miller, P.J., Niehuis, S., and Huston, T.L. 2006. "Positive Illusions in Marital Relationships: A 13-Year Longitudinal Study," *Personaltiy and Social Psychology Bulletin* 32:12, 1579-94.

Murray, S.L., Holmes, J.G., and Griffin, D.W. 1996. "The benefits of positive illusions: Idealization and the construction of satisfaction in close relationship," *Journal of Personality and Social Psychology* 70:1, 79-98.

Murray, S. and Vargas, M. 2020. "Vigilance and Control," *Philosophical Studies* 177, 825-843.

Moscovitch, M. 1989. "Confabulation and the frontal systems: Strategic versus Associative Retrieval in Neuropsychological Theories of Memory," In H.L. Roediger and F.I. Craik (eds.) *Varieties of Memory and Consciousness* (Hillsdale: Erlbaum), 133-60.

Newell, B.R. and Shanks, D.R. 2014. "Unconscious influences on decision making: a critical review," *Behavioral & Brain Sciences* 37:1, 1-19.

Nisbett, R.E. and Wilson, T.D. 1977. "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84:3, 231-259.

Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P.E. 2013. "Predicting ethnic and racial discrimination," *Journal of Personality and Social Psychology* 105:3, 171-92.

Penton-Voak, I.S., Rowe, A.C., and Williams, J. 2007. "Through rose-tinted glasses: Relationship satisfaction and representations of partners' facial attractiveness," *Journal of Evolutionary Psychology* 5:1, 169-81.

Ramachandran, V.S. 1996. "The Evolutionary Biology of Self-Deception, Laughter, Dreaming and Depression: Some Clues from Anosognosia," *Medical Hypotheses* 47, 347-62.

Ratcliff, R. and McKoon, G. 2008. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks," *Neural Computation* 20:4, 873-922.

Richardson, H. 1994. *Practical Reasoning about Final Ends* (Cambridge: Cambridge University Press).

Richardson, G., Dickinson, P., Burman, O.H.P., and Pike, T.W. 2018. "Unpredictable movement as an anti-predatory strategy," *Proceedings of the Royal Philosophical Society B* 285: 20181112. http://dx.doi.org/10.1098/rspb.2018.1112.

Robins, S. 2016. "Misremembering," *Philosophical Psychology* 29, 432-47.

Robins, S. 2020. "Mnemonic Confabulation," *Topoi* 39, 121-32.

Scanlon, T.M. 1998. *What We Owe to Each Other* (Cambridge, MA: The Belknap Press).

Schnider, A. 2017. *The Confabulating Mind* (Oxford: Oxford University Press).

Schnider, A., Treyer, V., and Buck, A. 2000. "Selection of currently relevant memories by the human posterior medial orbitofrontal cortex," *Journal of Neuroscience* 20:15, 5880-84.

Shepherd, J. 2015. "Consciousness, free will, and moral responsibility: Taking the folk seriously," *Philosophical Psychology* 28:7, 929-46.

Solomon, B.C. and Vazire, S. 2014. "You are so beautiful…to me: Seeing beyond biases and achieving accuracy in romantic relationships," *Journal of Personality and Social Psychology* 107:3, 516-28.

Stuss, D.T., Alexander, M.P., Lieberman, A., and Levine, H. 1978. "An extraordinary form of confabulation," *Neurology* 28:11, 1166-72.

Sullivan-Bissett, E. 2015. "Implicit Bias, confabulation, and Epistemic Innocence," *Consciousness and Cognition* 33, 548-560.

Swami, V., Furnham, A., Georgiades, C., and Pang, L. 2007. "Evaluating self and partner physical attractiveness," *Body Image* 4:1, 97-101.

Swann, W.B. and Read, S.J. 1981. "Self-verification processes: How we sustain our self-conceptions," *Journal of Experimental Social Psychology* 17, 351-72.

Tomasello, M. 2016. *A Natural History of Human Morality* (Cambridge, MA: Harvard University Press).

Tomasello, M., Melis, A.P., Tennie, C., Wyman, E., and Herrmann, E. 2012. "Two Key Steps in the Evolution of Human Cooperation: The Interdependence Hypothesis," *Current Anthropology* 53:6, 673-92.

Turner, M., Burgess, P., Cipolotti, L., Frith, C., Gilbert, S., Shallice, T., and Simons, J. 2008. "Confabulation, reality monitoring and the inferior medial PFC: Converging evidence from lesion and functional imaging studies," *Journal of the International Neuropsychological Society* 14, 67-68.

Turner, M. and Coltheart, M. 2010. "Confabulation and delusion: A common monitoring framework," *Cognitive Neuropsychiatry* 15:1-3, 346-376.

van Dellen, M.R., Shah, J.Y., Leander, N.P., Delose, J.E., and Bornstein, J.X. 2015. "In Good Company: Managing Interpersonal resources that support self-regulation," *Personality and Social Psychology Bulletin* 41, 869-882.

Vargas, M. 2013. "Situationism and Moral Responsibility: Free Will in Fragments," In T. Vierkant, J.Kiverstein, and A. Clark (eds.) *Decomposing the Will* (New York: Oxford University Press), 325-49.

Velleman, D. 2009. *How We Get Along* (Cambridge: Cambridge University Press).

Wheatley, T. 2009. "Everyday confabulation," In Hirstein (ed.) *Confabulation*.

Wilcox, W. and Marquardt, E. 2011. "How Parenthood Makes Life Meaningful and How Marriage Makes Parenthood Bearable," *National Marriage Project* (Charlottesvill, VA: Institute for American Values).

Wilson, T.D. and Nisbett, R.E. 1978. "The Accuracy of Verbal Reports about the Effects of Stimuli on Evaluations and Behavior," *Social Psychology* 41:2, 118-131.