CONFABULATING WELL
The Ethics of Confabulation

by

KATHLEEN MURPHY-HOLLIES

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Department of Philosophy
School of Theology, Philosophy and Religion
College of Arts and Law
University of Birmingham
February 2024

ABSTRACT

This PhD by papers is about the role of confabulation in our ethical and epistemic lives. Broadly, it is a defence of confabulation as much more than just irrationality, as a failure of human beings to accurately track evidence in identifying the reasons for their choices and behaviours. It is also a defence of confabulation as more than just a way of deceiving oneself about the true nature of one's moral character.

It is commonly stipulated that virtuous behaviour ought to be 'done for the right reason', but in cases of confabulation we may posit these 'right reasons' after the behaviour takes place. This is done in a completely earnest fashion with no desire to deceive. In these cases, individuals seem to pay ethically and epistemically, in that their behaviour was not in fact done 'for the right reason', and they don't become aware of that fact. However, this thesis argues that confabulation is not just an unwelcome blockage to enquiry in this way. In fact, it can be the beginning of enquiry, and it's a good thing that agents are moved by social pressures to formulate and share justifications of their actions. Where agents go wrong is not in confabulating, but in having poor skills of social negotiation afterwards. This is because understanding ourselves and the reasons for our own behaviour isn't actually an individual endeavour, but rather requires the input and appraisals of others. I describe the details of this in paper 1, and in paper 2 I describe the positive value of confabulations.

The remaining papers each apply this account of confabulation to a different context; political decision-making (paper 3), navigating upheaval and uncertainty (paper 4), conspiracist thinking (paper 5), and intimate relationships (paper 6). In essence, I argue that instead of agents trying somehow to never confabulate, they should aim to acquire certain prosocial skills and attitudes which will enable them to gain the benefits of confabulation, and therefore to confabulate well.

Acknowledgements

This PhD is for Andrew, who, while I was writing this, did everything else and more.

I did a lot and grew a lot as a PhD student, and it took an academic village. A lot of this village was Lisa Bortolotti, my primary supervisor. I will never be able to adequately express my gratitude for Lisa and everything she has done for me, and I'm very honoured that papers we wrote together form part of this PhD. Lisa made her philosophical wisdom, kindness and support feel inexhaustible to me, and it is one of my biggest ambitions to be able to embody this myself for others one day.

Discussions with my other supervisors, Iain Law and Quassim Cassam, hugely improved my papers and honed my philosophical aptitude. I will always be grateful for their input and spoiling me with warm, cooperative and rich philosophical exchanges.

The philosophy department at the University of Birmingham afforded me many opportunities to present my work and discuss philosophy fruitfully, and I am especially thankful for the meetings of the Women in Philosophy group.

My time working with Pilar Lopez-Cantero at Tilburg University was an absolute highlight of this PhD, and proved absolutely crucial for personal growth and the development of my capacities for emotional regulation. It was also a privilege to think, drink, talk and write philosophy with Stefaan Blancke during this time. I am grateful to the whole philosophy department at Tilburg University for being so friendly and welcoming during my research stay, which was joyful and exciting.

I am grateful to Christian Miller and the other participants at the Honesty Summer Seminar Series at Wake Forest University in the summer of 2023. Discussions had here with other early career researchers about honesty and virtue were great fun and hugely insightful.

I am indebted to many wonderful women who were immense pillars in my academic village, providing friendship and academic support: Ema, Katie, Chiara, Heather, Zuzanna, Areti, Noelia, Jess, Lucy, Francesca.

There were also kind and helpful men: James, Francesco, Joaquim, Matthew, Martin, Merten, Alf, Yannick.

Contents

Introduction....1
1. Scene Setting....1
2. Overall Significance of the project.... 2
3. Background on Confabulation....3
3.1. Clinical Confabulation.... 3
3.2. Confabulation and Motivation.... 6
3.3. Everyday Confabulation.... 9
4. Reasons and Virtue 10
4.1. Confabulation and Virtue.... 12
5. Overview of Papers.... 14
6. Appendix Material....17
7. Future Directions....18
Bibliography.... 20

Paper 1: The Know-How of Virtue.... 23
Paper 2: Stories as evidence....43
Paper 3: Self-Regulation and Political Confabulation....56
Paper 4: Exceptionalism at the time of COVID-19: where nationalism meets irrationality....75
Paper 5: Are Conspiracy Theorists Confabulating?....98
Paper 6: Confabulation and Reasons for Love....119

Appendix 1: Why We Should be Curious about Each Other.... 147
Appendix 2: What is left of irrationality?.... 160

Introduction

1. Scene Setting

Our sense of self is a central and transformative aspect of our ethical lives. If a concern with one's self is too central, we risk being egocentric and shallow. But if there is no concern at all with one's self, we risk being naïve and – somewhat paradoxically – self-absorbed, given a lack of regard for how we appear in the eyes of others.

Julia Annas suggests that virtuous behaviour ought to feel like playing the piano; smooth and confident. There is a sense of flow and engagement, in which thoughts of the self are far out of mind. Any sudden self-consciousness - is this really me? - seems to be something that would break this flow, and the deep engagement with the task at hand. But at other times, some self-consciousness is precisely what is called for in ethical behaviour. In response to abominable behaviour, we will often ask 'What kind of a person *does that*? They should take a long hard look at themselves'.

The centrality of character in virtue ethics is a reason that I chose to consider the ramifications of confabulation within this normative framework. Whether you are doing something 'for the right reason' is a stipulation of virtuous behaviour which particularly aims to capture that you are engaging with that behaviour 'with the right spirit', in a way which reflects a holistic goodness of character. You have been moved by a particular value, and now that value is at the forefront of your mind in guiding your behaviour. To be moved in such a way speaks to having an appropriate and heartfelt motivation to see goodness win out for its own sake.

Before starting this PhD, I had assumed that confabulation would only muddy the waters of acting for the right reason, in bringing us to simply posit the right reasons and good motivations after the fact, in a post-hoc way. I expected to outline ways to avoid or mitigate confabulating. However, I've come to see confabulation as a tool in managing this double-edged nature of self-awareness, with the relevant benefits and risks depending on how it is wielded and in what sorts of environments. Confabulation plays a role in both interrupting and in maintaining a smooth conception of the self, as coherent and consistent over time. This is because in confabulating, we have been prompted to pay closer attention to our behaviour, and we may face discrepancies between our account of that behaviour and the accounts that others give. But we also, in confabulation, explain ourselves in ways which protect our sense of self as consistent and positive, even when our insight is limited. This interruption and protection of one's sense of self can accord epistemic benefits to agents and contribute to the development of virtue. This PhD

attempts to spell out the mechanisms behind how this can take place, particularly as a dynamic process which feeds back into itself over time.

2. Overall significance of the project

I believe that taking a closer look at reason-giving which manifests individuals' imperfect and limited cognitive powers teaches us some valuable lessons about human beings.

The first is the extent to which people fill in the gaps in their understanding with details which make for a good fictionalisation of their lives. People's ambitions, desires, optimism and idealised selves are deeply woven into their self-interpretations and their decision-making. When people portray themselves in a flattering and self-serving light, it is tempting to hold this against them and find them self-aggrandising, even pompous. But this is hasty and, although will no doubt sometimes be an apt assessment, will not always be. There are good reasons to allow people to embellish their accounts, because they are not impotent embellishments. Under certain circumstances, they have the power to shape future behaviour, and they also tell us valuable information about that person.

Taking the forward-looking power and influence of these embellishments seriously means that we gain a more realistic account of how people develop and mature morally, which departs from accounts which posit ideal agents and unrealistic demands of them. It is too demanding to require that individuals already know whether a self-ascription is strictly accurate or not, and to only allow those positive self-ascriptions when all the evidence supports it. Given that real individuals have practical goals and limited cognitive resources, confabulation is a crutch which will often be used and therefore often be present in the development of virtue. I argue that virtue ascriptions can be appropriate and valuable for the development of virtue even if not all the traditional requirements for ascribing that virtue are in place yet.

We see attributions of virtue and vice be attributed to one another constantly. Yet, what is 'generous' to one person is 'patronising' to another. My account of confabulation illuminates the mechanism for how we socially calibrate our expectations of how to behaviourally manifest values, not only on an individual scale but also with regard to shared political and cultural values. Nowadays, we see that individuals are asked what is the 'feminist' answer to an issue, or what is the 'kindest' way to respond to a global crisis. My project therefore illuminates how individuals *come together* to answer these questions. The social dimension of virtue development has been recognised for some time, but my project really focuses on the mechanisms for how limited and imperfect agents use and benefit from each other in getting around their individual limitations. The mechanisms I describe also illuminate how the harm of prejudiced and harmful cultural

narratives filter down to affecting the lives of individuals and their interpretations of themselves and others.

I discuss confabulation in a number of broad interpersonal contexts, particularly political contexts. There is increased interest recently in the nature of political exchange, and particularly of political values, given issues of polarisation and misinformation. My analysis of imperfect reason-giving in these exchanges further illuminates the other motivations and practical concerns that individuals have in these contexts, and recognising these suggests certain ways of navigating these interactions. For instance, in discussions with others who have very different political ideals and beliefs to ourselves (such as conspiracy theorists), it is useful to be mindful of whether that conversation is about explaining the world or about individuals justifying themselves and meeting social goals. Locating the 'tone' of the disagreement here will highlight what kind of resolution is required. If the disagreement is rooted in competing explanations of the world, empirical research and evidence can help resolve the dispute. But if the disagreement is more so based in attempts to justify certain values, the resolution will look different. I return to these issues in section 1.6, 'future directions'.

I put forward an account of confabulation, as an inherently social and interpersonal phenomenon, and I emphasise its positive role in *constituting* self-knowledge. This is a radical departure from views which present confabulation as an individual deficit in rationality, and as a decisive threat to first-person authority and self-knowledge (Such as Scaife 2014 and Carruthers 2011). Where traditional approaches such as these would simply advocate for banishing such irrationality (or at least that this would be desireable), I describe exactly how agents should work *with* this reality of being human, as creatures who idealise and who are socially embedded. In doing so, I begin to tackle a very broad but central ethical question: How, as social agents with imperfect and limited faculties, we become good people and truly embody good traits and values.

3. Background on Confabulation

3.1 Clinical Confabulation

"Confabulation" derives from the Latin verb "fabulari" which means to tell a story, and "con" indicates that this story-telling is done together. "Fabula" is an old English word which is more recognisable as 'fable'. In essence, 'confabulation' describes the sharing of fables together (Hirstein 2005, 7). Here, I give a bit of background on history and use of the term 'confabulation', which has its roots in psychiatry where it was recognised as a symptom of a number of psychiatric disorders.

Uses of 'confabulation' in clinical settings are more likely to define confabulation based on its aetiological features, which centre the physiological mechanisms that underlie confabulation. Confabulation has been associated with memory impairment primarily, described as a tendency for patients to give false reports regarding memories instead of admitting not being able to remember something (Hirstein 2009, 3). Patients may not be able to retrieve memories (and so create false memories), or they may distort memories, make mistakes in establishing the temporal order of their memories, alter the context of memories, or cannot distinguish between what is a memory and what is made up by them (Bortolotti & Cox, 2009, 954). Fotopoulou suggests that it may be that the retrieval process is no longer guided by the appropriate processes which keep the agent engaged with the task at hand, and is instead led by associations which quickly disengage from the current context and question. For example, a patient might be asked what they were doing yesterday, and despite having been in the hospital all day, they will talk about having been at work, what their job requires of them, their performance at work etc. Temporal Context Deficits result in patients mistakenly thinking that past memories are relevant to the present moment, unable to 'turn them off' and thinking they should be, for example, at a meeting. Finally, Source Deficits mean that patients are unable to accurately source mental representations in their head as memory or dream, past or present, internal or external, making reality-monitoring difficult (Fotopoulou 2009, 266-9)

Confabulation is therefore seen, unsurprisingly, alongside memory disorders in particular, such as Korsakoff's syndrome, Alzheimer's disease, and aneurysm of the anterior communicating artery (ACoA) – which can lead to severe amnesia (Hirstein 2005, 8). Patients with Korsakoff's syndrome suffer chronic memory disorder, often due to alcohol misuse. These patients and those with amnesia will often confabulate when asked a question they cannot answer because of their extensive memory loss.

The word 'confabulation' was first used in a broader capacity beyond just memory deficits when used to describe patients who would provide answers to questions which they could not possibly know, instead of saying that they did not know or realising that they did not have the required knowledge. Norman Geschwind was the first to use the term 'confabulatory' in 1965 to describe participant responses like this, when they described stimulus which was not actually shown to them. He made no specific reference to memory processes (Hirstein 2009, 4). In another example, split-brain patients who have had the connections between their hemispheres severed, give explanations of their actions or choices despite their hemispheres being unable to communicate and produce a full, accurate answer (Hirstein 2005, 10). In essence, only the left hemisphere can formulate responses to questions, and it was found that patients' left hemisphere would produce plausible explanations for what the left side of the body is doing

(which it does not control), even though it cannot know why the left side of the body is acting as it, because it cannot communicate with the right hemisphere which controls it (Gazzaniga 1995). This suggests that an aetiological definition of confabulation may have to refer to a wider set of processes than just mnemonic ones.

Other conditions which do not centre around memory loss but which can still include confabulation include Anton's syndrome, Capgras syndrome, Schizophrenia and Anosognosia for hemiplegia (Hirstein 2005, 8). Patients with Anosognosia and with Anton's syndrome display a lack of insight into their illness. Patients with Anton's syndrome will answer questions about what they can see, giving plausible answer about, for example, what the room and the doctor look like. These patients, however, have cortical blindness (blindness which is due to brain damage rather than damage to any part of the eyes). Patients with anosognosia will give 'excuses' for being unable to move paralysed parts of the body, such as 'I'm too tired, doctor' or 'I just don't feel like it'. (ibid, 10-12).

Hirstein frames all clinical confabulation which is associated with some neural breakdown in a two-phase approach. Firstly, there is a deficit to some mnemonic or perceptual ability. Memory-based disorder can give rise to the former, and the separation of hemispheres can give rise to the latter. Secondly, according to Hirstein, there is also damage to prefrontal executive processes which, when functioning normally, would monitor self-generated explanations to check if they correspond with reality or face substantial counterevidence. This would ordinarily lead someone to realise that their explanations cannot be right, but patients do not seem to see that their explanations cannot be accurate (Hirstein 2009, 4). This 2-factor framework is akin to the two-factor accounts of similar phenomena such as delusions (Davies et al, 2001), but others posit only one-factor (Maher 1974; Gerrans 2002).

There are a few other distinctions in cases of clinical confabulation which are seen, and it is unclear whether they represent spectrums of severity of confabulation, or distinct types of confabulation. Features which we may distinguish across different types of confabulation may regard form or content. In terms of form, confabulation may be classed as spontaneous or provoked; it may be produced unprompted or it may be triggered in response to being questioned. Provoked confabulations are more likely to be momentary, whereas spontaneous confabulations are more likely be persistent, as there may be a consistent 'outpouring' of false memories etc (Bortolotti & Cox 2009, 953). With regards to content, confabulations may be fantastic or mundane; they may be particularly bizarre and imaginative, or they may include more everyday plausible memory distortions (ibid).

Some argue that spontaneous confabulation and provoked confabulation are completely separate disorders, as provoked confabulation occurs widely and is not associated with damage

to a specific brain region, whereas spontaneous confabulation is associated with lesions of the anterior limbic system and has more specific symptoms (Schnider 2004, 37-38). However, spontaneous confabulation can become provoked over time in patients who show improvement (Hirstein 2005, 9-10). Similarly, minor distortions can progress into more fantastical ones, which challenges any categorical divide between provoked/mundane/momentary confabulations and spontaneous/fantastic/persistent ones (Dalla Barba 1993).

It quickly becomes difficult for a clinical definition of confabulation to encompass the myriad of aetiologies and mechanisms which underlie instances of confabulation across many different psychiatric disorders. Some have attempted instead to define confabulation using its epistemic features, in an epistemic definition. A definition of this sort will focus on the epistemic features of confabulation; those which relate to knowledge and beliefs. These will be expressed in more easily recognisable surface features of confabulation, rather than focus on a link to a specific mechanism. Whereas a narrow definition of confabulation focuses on 'false memories' as the main feature of confabulation, a broad definition can encompass more types of false or unjustified belief which may or may not be about the patient's personal history or self-concepts (Bortolotti & Cox, 2009, 953). I return to definitions of confabulation which focus on epistemic features, and therefore are more easily applicable to everyday cases of confabulation, in subsection 3.3.

Some researchers claim that neither a wholly aetiological account of confabulation nor a wholly psychological / motivational one will be able to explain all the features of confabulation, and that therefore a hybrid account may be most successful (Fotopoulou 2007, 6). A role for motivation in clinical confabulation is more highly disputed, as some form of neurological damage is the key factor in causing confabulation which is not dependent on a patient's motivations. However, in cases of clinical confabulation following devastating brain injury, patients could be motivated to hold on to and defend optimistic beliefs concerning their health and their abilities despite them being profoundly ill-grounded. I explore these issues further in the next subsection.

3.2 Confabulation and motivation

Definitions which focus on the aetiology and mechanisms of confabulation, however much they may vary with respect to the specific brain lesions and corresponding disorders, tend to explain the negative features of confabulation much better than they explain the positive features. These positive features concern the often bizarre and imaginative nature of the confabulations, the specific content which they often focus on (repetitively), and the fact that these topics tend to be ones which have personal significance. They also tend to cast themselves and their situations

in a positive light (Turnbull et al 2004; Fotopoulou et al 2007; Fotopoulou et al 2004). If there is a retrieval deficit, why are certain memories consistently retrieved? And certain temporal and source misattributions consistently 'favoured'?

Turnbull and colleagues suggest that these features will require an acknowledgement of motivational and emotional factors in confabulation in order to explain them. They found that patients appeared to be *motivated* to inhibit reality-monitoring processes and that emotional states might influence exactly which ones (Turnbull et al 2004, 7). These reality-monitoring systems are what Hirstein describes as making up the 'second phase' of breakdown in cases of confabulation, where mnemonic and/or perceptual deficits make up the first.

Turnbull and colleagues describe a patient who mistakes a stranger for one of his close friends who used to live abroad with him, but who had in fact died many years earlier. Seeing the friend made the patient much happier, and despite acknowledging that "being dead in one country and alive in another may cause some legal difficulties", deficits in his executive processing and emotional inhibition meant that the possibility that the stranger was his friend was not moderated fully (2004, 13). This is despite being perfectly capable, cognitively, of evaluating some of the possible legal implications of being dead in only one country. This suggests an emotional motivation to only evaluate certain possibilities which occur to the patient. Turnbull and colleagues also found more widely that 80% of patients' false beliefs had a 'positive affect bias' and were more pleasant than the actual circumstances of the patients (ibid, 7). They suggest that this could mean that confabulation fulfils a protective 'mood-enhancing' function, and that perhaps the brain damage they suffer leaves their emotional systems intact but they lose the regulation and inhibition of those states and motivations with the loss of executive processes (ibid, 8, 13).

Similarly, Fotopoulou et al (2007) found a patient's confabulations to be mostly wishful, contain a lot of self-references, and be mostly self-enhancing. They protected what he saw as his positive features, all often in contexts in which these things were irrelevant (2007, 9, 11). They claim that an account of confabulation which only relates to cognitive impairment would not be able to explain this (ibid, 13-4). In healthy individuals, normal memory recall is affected by similar motivations and emotionally charged biases which determine which memories are recalled and how they are remembered. They are often remembered in such a way as to support certain self-narratives and fit into the wider scheme of an individual's goals and sense of purpose (Fotopoulou 2009, 278). These influences would normally be moderated and inhibited by executive processes, but following damage to these, they become too strong. This means that memory retrieval and executive monitoring processes become too heavily guided by motivations, biased by our concepts of our 'ideal' selves and not being checked against reality

and other more accurate memories (Fotopoulou 2009, 278-80; Fotopoulou 2007, 14). In essence, the suggestion is that the degree of damage to executive processes, which would normally control the competing cognitive and affective processes involved in memory and moderating what we say, explains the degree of reduced inhibition, awareness and self-monitoring in cases of spontaneous bizarre confabulations (Fotopoulou 2009, 265).

Taking a closer look at some cases of anosognosia reveals how patients may be motivated to preserve their self-concepts and their valued pre-morbid traits and capabilities. Many patients with anosognosia could understand and accept that they had the condition but did not seem to make realistic adjustments to certain beliefs given the reality of their condition, and the evidence of their lowered abilities. Aimola Davies (2009) makes use of the dual-process model of reasoning to describe how this happens in patients with anosognosia.

This model posits two distinctly different types of cognitive processing. These are 'system 1' and 'system 2' processes (Kahneman 2011). System 1 processes are driven by emotion, quick, compelling, make use of heuristics, take place automatically, do not demand many cognitive resources and tend to be inaccessible to introspection (Haidt 2001, 818 and 820). System 1 processes can give rise to gut feelings or intuitions which we may find persuasive despite not knowing their causes or the reasoning behind them (Railton 2014, 826-7). Only when we feel we need to, we start using system 2, which is a slower, effortful, explicit and sequential reasoning which is accessible to us and under our control (Haidt 2001, 818). Aimola Davies suggests that the more considered and intentional system 2 processes normally inhibit the knee-jerk reactions of system 1 but are not doing so properly in cases of anosognosia (2008, 192).

Less inhibited system 1 processes mean that a patient's belief that they will return to some valued activity, such as returning to work, is initially generated as a salient possibility (ibid, 194). Then, a motivational bias in using evidence to evaluate beliefs (ibid, 202) may explain why many patients were consistently unrealistic about their prospects of returning to work, despite being perfectly capable of understanding the consequences of their limitations when it came to other, less valued activities. One patient understood that he would be unable to play guitar anymore, but still thought that he could return to work (ibid, 215). As these patients demonstrate the cognitive ability to understand their diagnosis and its limiting effects on some activities, this suggests that it is motivation which affects their attention to and handling of evidence which doesn't support an ability to return to such valued activities. This could be because of the negative effects this would have on their valued self-representations. Emotionally-charged motivations to preserve positive self-representations would be based in system 1 processes, while executive processes and reality-monitoring would be more effortful system 2 processes. Normally, these would 'catch' any ideas (generated by system 1) which are strikingly unlikely to

be true despite being self-enhancing. But in anosognosia this does not happen. In these cases, system 2 processes do not sufficiently evaluate the belief that they will return to work given the evidence of their severe disabilities. The motivational bias is unchecked and continues to protect the belief that they will return to work, and the sense of agency encompassed by it.

There is still uncertainty as to whether motivation plays any role in patients beginning to confabulate in the first place or whether it only affects the content of confabulations, and also how exactly the interaction between cognition and motivation works (Blechner 2004, 18). But ultimately, I take motivational biases to explain why the content of confabulations tend to be positive and self-enhancing, indicating that some inferences are more likely to be made than others. I suggest that it is not difficult to see these same motivational factors present in everyday confabulation (although not to the same unmitigated extent when seen with significant neurological damage) when we evaluate some of our own self-concepts. I have already mentioned that motivations related to our desires, idealised selves and unfolding self-narratives are a normal part of memory retrieval, and it is not only in cases of psychiatric illness that the moderating influence of executive processes can fail to 'check' these influences. Older people also tend to be less inhibited (Fotopoulou 2009, 284) and even on an everyday level, we can notice perfectly cognisant individuals not updating some of their self-representations despite mounting evidence against them. Individuals may, for example, persistently overestimate their capacity for time-keeping. This gives a bit more detail to how motivations work in preserving positive self-representations, and given that believing ourselves to possess or be instantiating a certain virtue would be a positive self-representation, failing to accurately 'check' these self-representation with unbiased reviews of our own behaviour would be a concern for virtuous agents. Starting to consider whether we possess a certain virtue may risk triggering these kinds of motivations and our confabulations plaster over any evidence which does not support the self-ascription of that virtue.

3.3 Everyday confabulation

There is extensive debate regarding whether clinical and non-clinical instances of confabulation should be understood as being on one spectrum, differing in degree rather than kind, or as entirely separate phenomena. While some approaches (Hirstein 2005) embrace a move towards and a broad conception of confabulation which unifies clinical and non-clinical accounts, others argue for the two to be kept distinct. For example, Robins (2018) argues that while everyday confabulations can be understood as explanations or justifications, clinical mnemonic confabulations (which relate to inaccurate memories due to neurological damage) are better understood not in this way.

I am inclined to see clinical and everyday confabulation as differing only in severity rather than in kind, and can therefore be seen as being on a spectrum. As described above, one reason for this is the transferability, as I see it, of motivations involved in both clinical and everyday confabulation.

I outline key features of everyday confabulation throughout the PhD, and I lean very heavily on key features picked out by Ema Sullivan-Bissett (2015). These are that everyday confabulations are false or ill-grounded, provoked (this is far more common for everyday confabulations than clinical confabulations), motivated with regards to both confabulating in the first place and with regards to the specific content of those confabulations, they fill cognitive gaps, and finally confabulators have no intention to deceive. I think it is also useful to add that confabulations are post-hoc (Summers 2017) and that confabulators are confident in their confabulations (Hirstein 2005, 187). Arguably, too confident sometimes, to the extent that some confabulations might be described as resistant to counter-evidence (Bortolotti & Cox 2009, 956)

I also aim to emphasise, in my account of confabulation, how much of an interpersonal and unavoidable tool it is, employed (subconsciously) in agents' ongoing investigations into what drives their behaviour. Therefore, confabulation is not as inward looking as it initially seems, given its associations with deceiving oneself about less flattering portrayals of one's behaviour and moral character. It's a reflection and embodiment of our idealised selves, but I am keen to emphasise that these reflections are instrumental in bringing about those idealisations. Furthermore, given the cognitive limits in place with regards to individuals interpreting their own behaviour, it becomes one of the very few tools available for getting around those cognitive gaps. It's extremely valuable for cognition to not 'grind to halt' when the agent doesn't have all of the relevant facts of the matter totally understood and incorporated into her self-understanding, because this means that overall epistemic functioning is preserved. Solitary reflection and deliberation certainly play important roles in self-interpretation and self-understanding, but it is too demanding of everyday individuals for most self-knowledge to be gained in this way. Finally, I argue that the negative effects associated with confabulation should be more closely associated with the individual having antisocial attitudes which mean that confabulation can't be a useful social tool for them, rather than to confabulation itself being inherently bad.

4. Reasons and Virtue

Now that I have given some background to confabulation and what everyday confabulation looks like, I turn to thinking about confabulation in terms of virtue. In this subsection I spell out two prominent ways in which understanding the reasons for one's virtuous actions, choices,

attitudes etc. play an important role for the development of virtue. In a description of virtue given by Rosalind Hursthouse, we can pick out some of the key features of significantly cultivated virtue as (i) being a reliable disposition (of character) to (ii) act for certain reasons (iii) with the right accompanying emotional response and (iv) an acute perception of scenarios which encompass that virtue in some way, so that agents recognise the need for virtuous behaviour and are competent in situ (Hursthouse 1999, 11-12).

Firstly; reasons make the virtue. Hursthouse's description of virtue includes the criteria of acting 'for certain reasons', and notes that the virtuous agent ought to be able to give her reason for action and that this will illuminate to us not only why that action is appropriate in the eyes of that agent, but we learn about the character of that agent; what they have found salient, relevant, decisive, etc (1999, 123, 129). A more superficial or malevolent reason for action would clearly highlight that the behaviour in question falls short of virtue. Similarly, Aristotle's description of virtue as being moved "at the right times, with reference to the right objects, towards the right people, with the right motive, and in the right way" (Arist. NE II.6) also makes a direct reference to one's reasons and motivations needing to be 'the right ones'.

Secondly; Reasons enable agents to navigate varied and novel situations. Understanding one's reasons for some virtuous behaviour is what will enable agents to excel at recognising circumstances which encompass some virtue in some way, because they are well disposed to be motivated to act in relevant scenarios. If the reason you gave to a homeless person was to impress your date, you won't be motivated to help again when you are without a date. But if your reason for helping was compassion for that person, you will.

What is further required for excellence in virtue is what Annas calls a "drive to aspire" (2011, 16). Having this drive to aspire will mean that the budding virtuous agent will want to know *why* a certain response to certain situations is the right one, or in other words, what it is about the environment that *rightly* calls for that response and makes it *appropriate* (ibid, 23). True recognition of the call for some virtuous response is recognition that we have certain reasons for acting (Dent 1984, 23).

Without adequate accompanying understanding and wisdom, or 'phronesis', then this is an underdeveloped virtue akin to Aristotle's notion of 'natural virtue'; good behaviour which we are already inclined to do, but without some explicit insight into *why* it is so appropriate it can even be dangerous or harmful (Aristotle NE VI.13, 1144b). For example, a child may naturally have an inclination to share, but given a lack of insight into *why* it is good to share her toys then she will not grasp nuances. She may not grasp, for example, that she should also share her favourite toys and not just her less favoured ones.

So, developing virtuous behaviour entails identifying the reasons for certain responses, reflecting on them and over time, valuing them for their own sakes so that we can be independently motivated to be sensitive to those reasons in the future.

4.1 Confabulation and Virtue

Now that I have given the relevant background for my account of confabulation and on the value and importance of understanding one's reasons for the development of virtue, I briefly highlight a few ways in which I came to think that the two are likely to intersect. I structure these points around features of everyday confabulation which brought me to think that confabulation is indeed likely arise in the contexts of virtue development for individuals.

Confabulations are *provoked* (Sullivan-Bissett 2015, 551). Simply, We are very likely to 'be provoked', or asked about, virtuous behaviour, because our ethical behaviour is of very high concern to others and is central to social life in general. It is common that we are asked to explain or justify our ethical behaviours. Especially if we have behaved in some condemnable way, one of the first things we are asked to do is to 'explain ourselves'. Our ethical behaviour is so important, in fact, that I suggest we will often provoke *ourselves* for explanations of our behaviour. We find that individuals do in fact confabulate even when alone, or when no-one is going to hear their explanation (Wilson et al 1989, 297), which lends some support to this idea.

Secondly, confabulations are motivated. Sullivan-Bissett describes two key motivations in confabulation. Firstly, there is a motivation to give an answer at all in order to avoid the embarrassment of dumbfounding, and secondly, there is a motivation to give an answer which preserves positive self-representations (2015, 552). Being able to give an explanation of our *moral* behaviour is especially important, so the stakes are even higher to avoid dumbfounding. Particularly in the aftermath of condemned behaviour, one's reputation can be salvaged if we can 'explain ourselves'. So, we will be extremely motivated to give *an* explanation, even if we might not be in a position to give a particularly accurate one.

We will also want that explanation to protect positive self-representations. If we think of ourselves as generous or patient, we are keen to provide an explanation which supports and coheres well with that positive self-representation, rather than challenges it. The concept of a virtue is perhaps one of the most explicit and highly regarded forms of a positive representation, particularly as *morally good* representations. We will automatically want to protect the idea of having that virtue, that positive self-representation. So, it is more expectable that we will

confabulate reasons for our behaviour in instances in which perhaps we did *not* live up to some virtue.

Finally, everyday confabulations fill a gap (Sullivan-Bissett 2015, 552). More accurate explanations of behaviour may be *strictly unavailable* to us if they are completely opaque to introspection or impossible to retrieve (Sullivan-Bissett 2015, 554). Bortolotti highlights that agents often do not have introspective access to the *causes* of their attitudes and choices – though they have access to the attitudes or choices themselves (2018, 235). Alternatively, more accurate explanations might be *motivationally unavailable* to agents if they are not retrieved because the agent is not sufficiently motivated to retrieve them (Sullivan-Bissett 2015, 554). We are likely to find that explanations of our behaviour which feature our *vices* are going to be motivationally unavailable to us. We are instead motivated to preserve positive self-concepts, and we see evidence of this in common biases such as the better-than-average-effect, where individuals judge themselves as better than average in a number of domains, such as attractiveness, intelligence, and *moral character* (Brown 2012).

Ideally, virtuous agents act from a robust and adaptable moral intuition. It will have been honed through extensive experience and practice of virtuous behaviour, so that agents can 'see' intuitively how to appropriately respond to ethically charged scenarios, and are moved by the relevant *reasons*, manifested in those situations, for certain virtuous behaviour. As Annas describes, virtuous agents will be able to act quickly and fluently but will still be able to explain why some response is the best and appropriate one, not unlike an expert in some skill who can work quickly and swiftly but still mindfully (2011, 29). Aristotle too emphasised that no set of rules would be extensive enough; instead, agents will need a flexible faculty to guide them in situ, where the virtuous agent "feels and acts in accordance with the merits of the case" (Aristotle NE III.7, 1115b19-20).

Having moral intuitions is something familiar to most of us, wherein we can experience a 'gut feeling' that something is just not right, or alternatively, that something is the right thing to do despite extensive personal cost or inconvenience. Ideally, given enough time and experience, we develop 'gut feelings' which reliably guide us towards the right thing to do. Railton suggests that this type of intuition-based decision-making and behaviour, is realised by the affective system (2014, 840). However, a key feature of the affective system and affective processing is that it is opaque to introspection (ibid, 826). This is where we find a cognitive 'gap', in that the causes of these gut feelings are *strictly unavailable* to us, but the causes of these gut feelings are important for ascertaining whether some behaviour is virtuous, and for the agent to be aware of possible unwelcome factors influencing their behaviour. So, despite this requirement, the quick and

effortless 'evaluative perception' required of virtuous agents provides a 'gap' which confabulations could often fill.

5. Overview of Papers

The first two papers of my PhD lay down more theoretical work and ideas regarding my account of confabulation, and the remaining four papers are applications of this account to some distinct contexts, in order to flesh out the mechanism of exactly how confabulation can influence self-concepts and behaviour over time.

My first paper, **'The Know-How of Virtue'**, is published in the *Journal of Applied Philosophy*[1]. This paper encapsulates the core argument of my thesis, which is that confabulations can contribute to the development of virtue, under certain circumstances. In essence, confabulations can contribute to the development of virtue because the rosy and optimistic images of the self which they emphasise, at the expense of strict adherence to the facts of the matter, have forward-looking power in making that idealised self-image a reality. However, for this to happen, one needs certain self-relational skills and attitudes. I draw on the regulative mindshaping framework of how we understand ourselves and others, to flesh out how exactly these skills, which require 'know-how', mean that an individual's idealised self-ascriptions and actual behaviour can align over time.

This paper is where I lay out the details of what exactly confabulation is 'guilty' and 'innocent' of. I do not argue that confabulation is an unalloyed good thing that we should never worry about. I acknowledge that there are circumstances where confabulation will and does inhibit the development of virtue. These will be when (i) agents do not have the skills and attitudes described, which aid social negotiations and the aligning of idealised self-concepts and behaviour after confabulation, and when (ii) the agent is in an unideal social or cultural environment with regards to expressing values well through their behaviour. In these cases, confabulation can indeed continue to mask the reality of our poor behaviour from ourselves. However, this does not change that given the fact that confabulations fill cognitive gaps, the shared reason-giving of confabulation is the only possible way that agents *can* improve. This is because no amount of individual reflection fills a gap which the agent simply does not have introspective access to.

[1] Murphy-Hollies, K., (2023), 'The Know-How of Virtue', in *Journal of Applied Philosophy*, DOI: https://doi.org/10.1111/japp.12704]

My second paper is '**Stories as evidence',** which is co-authored with Lisa Bortolotti and published in *Memory, Mind and Media*[2].

This paper is about whether the stories that people tell can be taken as evidence for anything, given that people can be inaccurate, prone to memory distortions, and confabulate. This was a particularly pressing issue at the time (and still is) because of the importance and influence of stories being told about the coronavirus pandemic during lockdown. Another feature of this time was the amount of stories which were curated and told online, so we focus on this aspect too. We argue that even if stories cannot be taken as evidence for facts of the matter, they can be taken as evidence of someone's values, what they care about, and how that person sees themselves and their capabilities. This paper was my first attempt to push the *positives* that confabulation can capture about people, rather than seeing confabulation as primarily a deficit; a form of ignorance which means that the confabulator herself and those around her, only lose out.

With the machinery of my account of confabulation fully extrapolated, my next four papers are applications of my picture of confabulation to specific contexts.

My third paper is '**Self-Regulation and Political Confabulation'** and this is published in the *Royal Institute of Philosophy Supplements*[3].

This paper is an application of my account of confabulation to the political domain. People often confabulate reasons for their political behaviours, such as who they vote for and support. In a similar way to how confabulating about oneself can lead to incongruence between one's professed values and traits, and one's actual values and traits (what actually motivates someone, for instance), I show that confabulating about political choices can mean that there's incongruence between one's professed political values and the reality of the political choices they make. This gave me the opportunity to apply the skill of self-regulation to the phenomenon of undermining propaganda, which has the same structure of one value being professed while another is really being furthered. I also describe in more detail how this skill is a virtue, whereas my "The Know-How of Virtue" paper describes how self-regulation is a skill.

---

[2] Murphy-Hollies, K. and Bortolotti, L. (2021), 'Stories as Evidence', in *Memory, Mind & Media*, 1(3). DOI: https://doi.org/10.1017/mem.2021.5.
We worked together at length to figure out exactly what we wanted to say with regards to what stories can still be evidence of, given the prevalence of confabulation, and this is the heart of the paper. We also thought of and sourced the numerous examples used throughout the paper, together. When it came to writing, Professor Bortolotti took the lead on writing the earlier sections which highlight some key features of stories and digital storytelling. I took the lead on writing the latter sections on confabulation and what confabulations can be evidence of.

[3] Murphy-Hollies, K., 2022. Self-Regulation and Political Confabulation. Royal Institute of Philosophy Supplements, 92, pp.111-128, DOI: https://doi.org/10.1017/S1358246122000170.

My fourth paper is '**Exceptionalism at the time of COVID-19: where nationalism meets irrationality'**, which is co-authored with Lisa Bortolotti and published in *Danish Yearbook of Philosophy*, in a special issue on 'Nationalism and Irrationality'[4].

This paper discusses the role of irrationality and confabulation during times of collective stress, uncertainty, and upheaval, such as the covid-19 pandemic and lockdowns. It is application of my account of confabulation to the decision-making and behaviour of individuals in these contexts, and I argue that confabulation played a role in justifying poor behaviour during these times – in particular, breaking covid-rules. This is therefore an example of confabulation justifying, and therefore perhaps prolonging, *bad* behaviour. This paper gave me the opportunity to think more about the interplay between individual narratives for behaviour and collective narratives – particularly, nationalist narratives of superiority. I argue that unrealistic optimism, despite being a likely contributor to behaviour which flouted the rules, is a factor that was overlooked in these confabulations.

My fifth paper is **'Are Conspiracy Theorists Confabulating?'**. This paper is currently under review at Review of Philosophy and Psychology.

In this paper I argue that understanding many conspiracy theorist claims as confabulations helps us make sense of them. Familiar motivations to justify oneself and protect our self-concepts can explain some of the most bizarre features of conspiracist claims, such as the tendency to endorse contradictory theories. Writing this paper therefore gave me the opportunity to see 'how far' I can push the everyday cognitive architecture of confabulation in explaining beliefs which strike others as markedly irrational but also, very often, ethically unacceptable. This paper also made me think further about confabulations about the self (i.e., how much one knows and how one knows it) and confabulations about the world, and how the two can relate to each other.

[4] Bortolotti, L. and Murphy-Hollies, K. (2022), 'Exceptionalism at the time of COVID-19: where nationalism meets irrationality', in the *Danish Yearbook of Philosophy*, 55(2), pp.90-111. DOI: https://doi.org/10.1163/24689300-bja10025
Lisa Bortolotti and I worked together in sourcing and discussing the numerous examples throughout the paper. Lisa took the lead in sections discussing optimism, and I took the lead in sections discussing confabulation and its role in justifying behaviour.

My sixth paper is **'Confabulation and Reasons for Love'**, which is co-authored with Pilar Lopez-Cantero[5].

In this paper, we argue that confabulation can play a positive role in the development of loving relationships, particularly in ways which speak to the 'normative reasons' for love posited by rationalists about love. It is an application of my account of confabulation to a complex and intimate interpersonal domain. The literature on reasons for love is split between those who think that there can be reasons for loving someone in particular, and those who think that there cannot be reasons for love. We add a whole new dimension to this literature by considering what it means when real people, as limited agents, confabulate reasons for why they love someone. We argue that confabulated reasons can still be 'legitimate' reasons for love for both camps in the literature, hence blurring the long-standing distinction between the two. Writing this paper gave me the opportunity to think more carefully about the nature of true but ill-grounded confabulations in particular, and the possibilities they offer in influencing behaviour.

6. Appendix material

The first paper included in my appendix is '**Why We Should be Curious about Each Other'**, co-authored with Lisa Bortolotti and published in *Philosophies*, in a special issue 'Between Virtue and Epistemology'[6].

In this paper Lisa Bortolotti and I focus in particular on the value of curiosity as an epistemic virtue, but also as a moral virtue. We draw attention particularly to how it is difficult to tease apart the benefits which the trait of curiosity brings to one's epistemic functioning and one's moral character. A context which really brings this overlap out is that of clinicians listening to their patients when they describe their struggles with mental health. We argue that curiosity helps mitigate epistemic injustice, because the capacities for self-knowledge and understanding of the patient are given full credibility and uptake, and at the same time, the patient herself is

[5] This paper was written during my junior visiting fellowship at Tilburg University, from mid-October to mid-December 2023. The division of labour for this paper was 50/50. Pilar is an expert in the philosophy of love and we combined this with my expertise in confabulation and the psychological research on introspection and reason giving to come up with this novel argument. Pilar took the lead in writing sections 1 and 5, which describe the problem of epistemic access and the relevance of our argument for anti-rationalists, respectively. I took the lead in writing sections 2 and 6, where I explain what confabulated reasons for love look like and whether confabulating about love is a good thing for relationships, respectively. We both wrote section 4, the heart of the paper, together. This section describes exactly how confabulated reasons for love can form the normative reasons for love which philosophers of love are interested in. We are imminently sending this paper to *Philosophers' Imprint*.

[6] Bortolotti, L. and Murphy-Hollies, K., (2023), 'Why We Should Be Curious about Each Other', in *Philosophies*, 8(4), p.71. DOI: https://doi.org/10.3390/philosophies8040071 . My contribution to this paper was 50%.

afforded full moral recognition and respect as an agent. Writing this paper was an opportunity for me to focus in more detail on a particular trait and its dual-nature in benefiting both individuals' epistemic and moral characters, as this is something which I think is the case for the skills and attitudes I posit as enabling individuals to get the best out of confabulating.

The second paper included in my appendix is a book review of Neil Levy's 'Bad Beliefs', co-authored with Chiara Caporuscio, and entitled **'What is left of irrationality?**[7].

In this review, Chiara Caporuscio and I review Neil Levy's arguments about the rationality of 'bad beliefs'. Bad beliefs are those which conflict with the beliefs held by relevant authorities on the matter, and which conflict with widely-known and easily available public evidence on the issue. Levy argues that many attempts to inculcate individuals against bad beliefs are too individualistic, when the problem really lies in one's environment. This is why he argues against virtue epistemology as a useful tool against bad beliefs. However, writing this review gave me the opportunity to push back on this idea and argue that epistemic virtues are actually very valuable in another practice which Levy *does* acknowledge as important and valuable to navigating bad beliefs; deferring to others. We argue against Levy's overly individualistic account of epistemic virtue, and instead emphasise their role in facilitating the *sharing* and exchange of knowledge between people. Again, these ideas complement my account of certain skills and attitudes contributing to agents' epistemic functioning, particularly in the aftermath of confabulating.

7. Future Directions

There are two main research areas I aim to pursue after completing this PhD.

The first is to explore the notion of group confabulation, as opposed to confabulation by individuals. Certain societies and cultures, as an individual unit or group, can overlook the role of certain historical factors in causing other key historic events, as well as in forming present hallmarks of that society (such as rituals, traditions, trademarks etc.). A similar story may apply to institutions, where companies and organisations prioritise certain narratives over others in giving reasons for certain business or policy decisions, with the aim of preserving a certain (collective) identity rather than of adhering to the evidence and facts of the matter. My final paper of this PhD, 'Confabulation and Reasons for Love' is a first foray into this research interest of narratives which belong to and describe a group, rather than an individual.

[7] Murphy-Hollies, K. and Caporuscio, C., (2023), 'What is left of irrationality?', in *Philosophical Psychology*, 36(4), pp.808-818. DOI: https://doi.org/10.1080/09515089.2023.2186220 . My contribution to this paper was 50%.

The second is to consider whether the harms incurred when confabulations are completely dismissed in their entirety, given the associated opportunities which are then lost for agents and the social goals which then go unmet, may be captured within the notion of *epistemic injustice*. All the contexts I investigate in this PhD (moral reasoning, political decision-making, adopting conspiracy theories etc) revolve around social exchanges of information, and epistemic injustice can surely occur. On my account, knowledge is not only transferred in these contexts but *created*, as agents employ and practice skills in social negotiation and self-understanding. Therefore, it would be fruitful to consider whether people's capacities as knowers (particularly in capacities for know-how and not only knowers of propositional knowledge) are undermined when confabulations are totally dismissed. People's senses of agency can be diminished here and this is familiar to cases of epistemic injustice. Individuals have to practise and posit agency in order to develop it and become able to embody it more so than they currently do. It is also plausible that certain groups are more likely to rely on these proleptic tactics for agency cultivation; individuals who have experienced upheaval and trauma, and/or members of marginalised and oppressed groups who do not have rich hermeneutical resources available to them in their cultures which emphasise their capacities for agency.

There are other related research topics here which I don't discuss in this PhD, but would be interested to see further developed. One is the details of the notion of 'ill-grounded', and whether one ought to favour an internalist or externalist standard for understanding this (poor) adherence to evidence. In this PhD I assume rather than argue for clinical and everyday confabulation being of the same kind but differing in severity, and I do not critically engage with Hirstein's two-factor account of confabulation. However, it would be interesting to investigate where exactly the difference between the clinical and non-clinical confabulation lies if we accept that both revolve around the rationalisation of feelings. When exactly (i) feelings and (ii) rationalisations rise to the level of being 'clinically abnormal' would be interesting to decipher. The feelings in question would be assumed to be less bizarre and strong in cases of everyday confabulation, but this might be hasty. Individuals' feelings that they make choices in response to good evidence and accurate appraisals of the world, and that their capacities for agency are robust, may be stronger and more resistant to counterevidence than we ordinarily assume. Finally, I assume doxasticism about confabulations; that people believe their confabulations. This claim is most clearly contestable when I argue that conspiracist claims can be understood as confabulations, because many suggest that conspiracist claims are not always sincerely believed. Whilst still holding onto the notion that confabulations at their core are beliefs, it would be interesting to consider whether other non-doxastic capacities such as imagination are involved in some way with confabulation. For example, forms of aesthetic engagement with the narratives of conspiracy theories may have a parallel in generating confabulations.

Finally, this PhD focuses on what the confabulator should be like in order to get the best outcomes of confabulating. But there is surely plenty to say about the best ways to respond to and receive reasons given by others, perhaps especially if they are particularly bizarre or implausible. Progress here will be greatly informed by empirical work as it develops. At the moment, I am interested in work here by Alessandra Tanesini and colleagues on humility and 'value affirmation'. They have found that having humility and taking time to affirm one's values before having conversations on highly polarised topics helps those conversations go well (Hanel, Tanesini and Maio, 2023). This empirical work provides some insight into best ways to navigate highly polarised topics and conversations in a way which works *with* the less epistemic and more social motivations of people, which complements my arguments.

Bibliography:

Aimola Davies, A.M.., Davies, M., Ogden, J.A., Smithson, M. and White, R.C., (2009), 'Cognitive and motivational factors in anosognosia', in Bayne, T. and Fernández, J., (eds) (2010), *Delusion and self-deception: Affective and motivational influences on belief formation*, Psychology Press, UK/USA.

Annas, J., (2011), *Intelligent Virtue*, USA, Oxford University Press.

Aristotle., (2009), *The Nicomachean Ethics*, Great Britain, Oxford University Press.

Blechner, M.J., (2004), 'Commentary on "The Pleasantness of False beliefs"', in *Neuropsychoanalysis*, 6(1), pp.16-20.

Bortolotti, L., (2018), 'Stranger than fiction: costs and benefits of everyday confabulation', in *Review of philosophy and psychology*, 9(2), pp.227-249.

Bortolotti, L. and Cox, R.E. (2009), 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, 18(4), pp.952-965.

Brown, J. D., (2012). Understanding the Better Than Average Effect: Motives (Still) Matter. *Personality and Social Psychology Bulletin* 38(2):209–219

Carruthers P (2011) The opacity of mind. Oxford University Press, Oxford

Dalla Barba, G. (1993). Different patterns of confabulations. *Cortex*, 29, 567–581.

Davies, M., Coltheart, M., Langdon, R. and Breen, N., 2001. "Monothematic delusions: Towards a two- factor account," *Philosophy, Psychiatry and Psychology*, 8 (2/3), 133–158.

Dent, N.J.H., (1984), *The moral psychology of the virtues*, Great Britain, Cambridge University Press.

Fotopoulou, A., Solms, M. and Turnbull, O., (2004), 'Wishful reality distortions in confabulation: A case report', in *Neuropsychologia*, 42(6), 727-744.

Fotopoulou, A., Conway, M., Griffiths, P., Birchall, D. and Tyrer, S., (2007), 'Self-enhancing confabulation: Revisiting the motivational hypothesis', in *Neurocase*, 13(1), 6-15.

Fotopoulou, A., (2009), 'Disentangling the motivational theories of confabulation', in Hirstein W., (eds) (2009), *Confabulation: View from neuroscience, psychiatry, psychology and philosophy*, Oxford University Press, USA.

Gazzaniga, M. S., (1995). Principles of human brain organization derived from split-brain studies. *Neuron* 14: 217-228.

Gerrans, P., 2002, "A one-stage explanation of the Cotard delusion," in *Philosophy, Psychiatry, & Psychology*, 9(1), 47–53.

Haidt, J. (2001) 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement', *Psychological Review*, 108(4), pp. 814–834.

Hanel, Tanesini and Maio, (2023), 'Want to avoid heated arguments? Try this technique before having a difficult conversation', in *The Conversation*, Accessed 19/02/2014. https://theconversation.com/want-to-avoid-heated-arguments-try-this-technique-before-having-a-difficult-conversation-199033

Hirstein, W., 2005. *Brain fiction: Self-deception and the riddle of confabulation*. Mit Press.

Hirstein, W. ed., 2009. *Confabulation: Views from neuroscience, psychiatry, psychology and philosophy*. Oxford University Press, USA.

Hursthouse, R., (1999), *On Virtue Ethics*, US, Oxford University Press.

Kahneman, D., (2011), *Thinking, Fast and Slow*, Great Britain, Penguin Books Ltd.

Maher, B.A., (1974), "Delusional thinking and perceptual disorder," in *Journal of Individual Psychology*, 30(1), 98–113

Railton, P., (2014), "The affective dog and its rational tale: Intuition and Attunement", in *Ethics*, no. 4 vol. 124, pp 813-859

Robins, S., (2018), "Mnemonic Confabulation", in *Topoi*, pp. 1-12.

Scaife R (2014). A problem for self-knowledge: the implications of taking confabulation seriously. *Acta Analytica* 29:469–485

Schnider, A., (2004)., 'Commentary on "The Pleasantness of False Beliefs"', in *Neuropsychoanalysis*, 6(1), 37-39.

Sullivan-Bissett, E., (2015), Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition*, 33, 548-560.

Turnbull, O.H., Jenkins, S. and Rowley, M.L., (2004), 'The pleasantness of false beliefs: An emotion-based account of confabulation', in *Neuropsychoanalysis*, 6(1), pp.5-16.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J., (1989), "Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do.", in *Advances in Experimental Social Psychology*, 22, 287–343

Paper 1: The Know-How of Virtue

Murphy-Hollies, K., (2023), 'The Know-How of Virtue', in Journal of Applied Philosophy. DOI: https://doi.org/10.1111/japp.12704

Journal of
APPLIED PHILOSOPHY

Journal of Applied Philosophy
doi: 10.1111/japp.12704

The Know-How of Virtue

KATHLEEN MURPHY-HOLLIES

ABSTRACT *It is widely accepted that virtuous behaviour ought to be motivated in the right way, done for the right reasons, and an appropriate response to the values manifested in a situation. In this article I describe how cases of individuals having poor understanding of the reasons for their behaviour, can nevertheless be conducive to the development of virtue. One way in which giving reasons for one's own behaviour can be inaccurate is when the reasons given are confabulatory. In confabulation, the reasons given for behaviour are* post hoc*, do not capture factors which actually brought the behaviour in question about, and are not well supported by evidence. Confabulations protect an individual's positive self-representations at the expense of more accurate appraisals of themselves and the circumstances. However, I argue that engaging in the construction of a positive self-narrative can be efficacious in making it a reality. Importantly, this is only possible when agents have the know-how of certain self-relational skills and attitudes, which are captured in a meta-virtue of self-regulation. When individuals can regulate their values and their behaviour effectively, such that they are in keeping, they can ultimately consistently embody virtues. Confabulation can be, and probably often is, part of this process.*

1. Introduction

It is widely accepted that virtuous behaviour ought to be motivated in the right way, done for the right reasons, and an appropriate *response* to the relevant values manifested in a situation. In this article I describe how cases of individuals having poor understanding of the reasons for their behaviour, can nevertheless be conducive to the development of virtue within them. One way in which identifying and giving reasons for one's own behaviour can be poor or inaccurate is when the reasons given are confabulatory. In cases of confabulation, individuals give reasons for their behaviour in a *post hoc* fashion. The reasons given do not accurately track the factors which actually brought the behaviour in question about, and they are not well supported by evidence.[1] Confabulations protect an individual's positive self-representations at the expense of more accurate appraisals of themselves and the circumstances they were in. However, I argue that engaging in the construction of a positive self-narrative can actually be efficacious in making it a reality. Importantly, this is only possible when agents have the know-how of certain self-relational skills and attitudes, which are captured in a meta-virtue of self-regulation. The self-regulation required, therefore, is between the positive self-representations which the individual holds, and the individual's actual behaviour, so that these things are in keeping. Individuals can then effectively regulate their values and their behaviour, such that ultimately they can consistently embody virtues. Confabulation can be, and probably often is, part of this process.

In Section 2, I outline what confabulation is and the worry it poses for the feasibility of virtuous behaviour. In Section 3, I outline the possible benefits of confabulation found in

Bortolotti's work.[2] In Section 4, I describe how these could be applied in the development of virtue. In Sections 4.1 and 4.2, I explain how the skilful meta-virtue of self-regulation enables agents to effectively gain and implement these benefits in aligning their behaviour with virtuous self-concepts. In particular, I describe how this self-regulation calls for a skill-based 'know-how' type of knowledge about the self, rather than merely propositional knowledge about the self. This is in line with the processes of mindshaping as described by McGeer.[3] In Section 5, I consider an issue which might arise from these processes taking place in unideal cultures.

2. Confabulation and the Challenge It Poses

In this section, I give an overview of features of confabulation which are discussed in the literature. I will then go on to describe why, given some of these features, confabulation gives us cause to worry about the feasibility of virtuous behaviour – particularly given the stipulation that virtuous behaviour be done 'for the right reason'. Some have argued that the prevalence of confabulation shows that we are much worse at knowing and understanding ourselves than we commonly think, and that we might therefore not have any more authority or ability to know ourselves than others do, through their third-person perspective.[4] This article joins others (such as de Bruin and Strijbos, and Andreotta)[5] in pushing back against this interpretation of confabulation research as a radical threat to our self-knowledge. In turn, this means that it is not (necessarily) a threat to the feasibility of truly virtuous behaviour.

Everyday confabulation has been studied in the contexts of consumer choice,[6] moral convictions,[7] and aesthetic preferences.[8] In all of these cases, individuals give an explanation or justification for a choice or action which is ill-grounded; that explanation is not based on the facts of the matter about what brought that behaviour about.[9] So, it does not capture all the relevantly efficacious factors in play, it makes inaccurate claims about the circumstances, and is not well supported by evidence.[10] For example, in what is now one of the most famous studies on confabulation, Nisbett and Wilson's stocking study, participants were asked to pick a favourite pair of stockings from a line-up and then explain their choice.[11] As often happens in justifying consumer choices, the participants would say that the particular item they had chosen was of especially high quality in some respect (such as being the softest or brightest, etc.). In reality, however, all the products presented to the participants were identical. This explanation, therefore, is very poorly supported by the evidence because the chosen stockings were in fact no different from the others. So, this cannot be what brought their choice about and leaves another factor, which does explain why they chose as they did, unidentified. In fact, stockings on the right-hand side were chosen significantly more often by participants due to a right-hand bias. However, this reason for their choice was never identified by any of the participants and they instead confabulated reasons about their particular choice of stocking being superior in some way.[12] Throughout this article, I will often refer to 'inaccurate' or 'unwarranted self-ascriptions', by which I mean these types of mistaken and poorly supported attributions of some state or motive to oneself in a confabulation.

Very often, these unacknowledged causal influences on behaviour would in fact not be endorsed by the individual if they were to become aware of them. This is because, as we shall see, a key motivation of confabulation is protecting one's rational reputation and

positive self-concepts,[13] and the unacknowledged causal influences rarely support these. The reality of their influence may even be quite harmful and unethical – even from the point of view of the agent herself. A prime example of this is behaviour that is caused by implicit bias. Imagine, for example, a hiring manager who chooses not to appoint a woman to the position of maths teacher, and then confabulates reasons for this choice afterwards when asked by others to explain it. Perhaps he says that her CV was inferior, when in reality her CV was identical to the CV of the man appointed instead (as has been found to happen).[14] In this case, the employer has a perfectly good grasp of what he is taking to have been *his own* reasons for action, otherwise known as his *motivating* reasons. However, he does not capture some importantly efficacious *causal reasons* for his behaviour, namely implicit bias.[15] The role that implicit bias played in his hiring decision does not support his basic self-conceptions of being a rational and competent hirer, and his more abstract and moral self-conceptions of being egalitarian. This example also demonstrates that the causal reasons for behaviour are not only important because of their wider ethical consequences, but also because they can impede our behaviour effectively realising our consciously endorsed values. They are therefore also important and relevant for truly virtuous behaviour which is consistent.

This example also shows that confabulations fill a cognitive gap,[16] which is why the employer was unable to identify implicit bias as a factor in his decision. Although the employer may have what Holroyd[17] calls 'inferential awareness' about the existence of implicit biases, and even that he probably has some himself which affect his behaviour at times, he cannot know through straightforward introspection at this time that his decision here was caused by implicit bias.[18] If we had this kind of introspective awareness, implicit bias would not be so infamously difficult to tackle, with solutions tending to lean on long-term, indirect self-management. In the same way, the influence that the positioning of the stockings had on the choices made by the participants in Nisbett and Wilson's study[19] was not introspectively accessible to them, and also demonstrates how confabulation fills cognitive gaps.[20]

All of this means that the employer cannot know that his explanation is ill-grounded and this is always the case with confabulation. When individuals confabulate, they completely endorse the account they are giving of the reasons for their behaviour. They believe the explanation to be accurate and they have no intention to deceive at all, they just do not have introspective access to more accurate insights about what drove their behaviour. This is the crux and enigma of confabulation; that instead of an individual's attention being drawn precisely to their lack of answer for the cause of their behaviour, they confidently fabricate reasons for that behaviour and completely believe them.

This means that when an individual confabulates, the only other possible alternative which we could have seen from them (of course, the individual never consciously chooses to confabulate or to do otherwise) is being uncomfortably dumbfounded and finding themselves not able to give an answer at all. We see that this is what eventually happens to individuals in Haidt's study, in which participants are asked to explain their ethical stance that a case of a brother and sister sleeping together is wrong.[21] Participants justify their stance with poor reasons which rest on misconstruing the specifics of the case, and when each one of their offered reasons is dispelled as ill-grounded, participants eventually end up uncomfortably dumbfounded. They are unable to give better reasons for their stance but remain unpersuaded to give it up. Possible benefits of confabulation, then, should be assessed in comparison with the agent instead being

dumbfounded, rather than with them having an accurate answer, because that option just is not on the table.

Many of these features make confabulations a particularly pernicious form of inaccurate reason-giving, and therefore a particularly pernicious obstacle to developing virtue. Confabulation is not just inauthentic *post hoc* rationalisation or bullshitting, in which agents do have cognitive access to the influence of these other causes on their behaviour but instead opt for more flattering portrayals of it. Confabulators are totally authentic and earnest in giving their account, and yet it is so starkly poorly supported by evidence and the facts of the matter – the CVs and the stockings are all in fact identical.

It is hopefully becoming clear why this poses such a worry for the feasibility of virtuous behaviour, which relies on an astute appraisal of the circumstances and their role in bringing behaviour about. Confabulation looks like a way to twist any previous poor behaviour into something much more flattering, without the agent even realising. Our employer thinks he is motivated by a concern to hire the best candidate for the job, but he instead perpetuates sexism. When his attention is drawn to his decision, instead of getting closer scrutiny, we get a confabulatory story about his decision which protects his positive self-representations.

What exactly is expected of virtuous agents when they act 'for the right reason'? Wanting behaviour to be done 'for the right reason' is wanting it to be done 'in the right spirit'.[22] The specific role and meaning of acting 'for the right reason' varies but is nevertheless emphasised by most virtue ethicists, beginning with Aristotle defining virtue as being moved 'with the right motive, and in the right way'.[23] Knowing *why* one is doing as one does is what develops the raw, natural virtue of children into mature virtue, where the agent understands *why* the behaviour is appropriate.[24] Hursthouse follows Aristotle in stating that virtuous agents ought to 'know what they are doing'[25] and have a reliable disposition to act for certain sorts of reasons.[26] What might these reasons look like? These reasons ought to refer to the value of the virtue itself, in some way. So, the agent recognises the value in 'honesty' or 'generosity', and that is reflected in the reason they give for acting as they did.[27] In other words, the virtuous agent chooses the virtuous action *for its own sake*, or because it is the honest thing to do, or the generous thing to do, and so on (allowing for wide variation in exactly how different agents get at these specific ideas). The reasons given by agents should tell us something about how they viewed the situation, how that brought them to see the action as good and appropriate, and the salience of the moral values reflected in it.[28] As we have seen, however, confabulation may piece together this kind of story after the fact, when we need it to drive the behaviour beforehand.

The requirement to act for the right reason disqualifies good behaviour which happens by fluke from counting as virtuous, and also good behaviour which is done for shallow reasons such as impressing someone. Annas describes further that acting for the right reason is what enables virtuous agents to be adaptive and flexible. For instance, to recognise the need for virtuous behaviour such as generosity across many circumstances where the specific environment will vary greatly, but the core moral impetus will not.[29] Although the reason an agent gives for an action may be a normative reason for that behaviour, that is, it does speak positively in favour of acting in that way, if the action does not actually successfully further that value then it fails to be virtuous. As Swanton describes, virtuous behaviour needs to 'hit the target'.[30] Although hiring the candidate with the strongest CV is a strong normative reason for hiring someone, our employer fails to actually do so (to hire the strongest candidate and embody his egalitarian values).

So, in throwing a spanner into the works not only of grasping the reasons for one's own action, but also in one's ability to realise this failure, is confabulation purely foe for the virtuous agent?

3. Benefits of Confabulation: Agency, Self-Image, and Social Engagement

Important motivations are fulfilled by confabulating. The employer does not see himself as having sexist attitudes, and so he preserves this positive self-representation with his confabulation.[31] Despite agents not having access to more accurate and well-grounded explanations for their behaviour, confabulation enables them to give an answer and avoid the embarrassment of being dumbfounded about their own behaviour.[32] This in turn enables them to signal to others that they are rational, competent, and trustworthy.[33] Others have described confabulations as having a narrative form, in which individuals organise their experiences and behaviour into a meaningful narrative.[34] This is not only very valuable for themselves to do, but it can also facilitate the communication and sharing of important values.[35] Finally, confabulatory explanations can provide individuals with a causal understanding of their behaviour and their circumstances.[36] In this section, I give more detail of the benefits which can follow from fulfilling these motivations and confabulating, not only epistemically but also – in the next section – morally, for the development of virtue. However, later on I outline what else individuals need to have in order to actually gain these benefits, given that they are far from guaranteed.

I divide the potential benefits of confabulation into roughly three areas: enhancing perceived agency, protecting a positive self-image, and enabling social engagement.

How does confabulation enhance a sense of agency? Confabulation constructs or enhances an agent's sense of acting for good, sensible reasons, and that these underlie their behaviour and choices[37] rather than overlooked factors which are not strictly relevant or rational. Then, these more sensible reasons can be grouped together into common themes, which reflect an individual's ongoing beliefs, motivations, and traits.[38] Thus, confabulation brings together and integrates self-related knowledge. By aligning behaviour with valued self-conceptions, for example, such as by reasoning that I spent time helping my friend the other day because I am a helpful person, agents are able to make sense of their past behaviour and predict their future behaviour.[39] When confabulations emphasise a certain trait which agents perceive themselves as having, this can provide a sort of blueprint for navigating future choices and behaviour by bringing agents to act in ways which are in keeping with that self-ascription.

Velleman suggests that this self-consistency is fundamental to agents being able to function and make any decisions at all.[40] The imposition of consistency and a sense-making narrative onto behaviour is the basis for how individuals consider their reasons for various actions and make decisions; they choose the option which makes the most sense and coheres best with an ongoing narrative they have of themselves. Thus, in deciding what action to take next, agents consider their own self-concepts, and do what makes sense or is a coherent and meaningful next step, in the light of them. According to Velleman, this process *is* deciding for reasons; 'I suggest that the narrative background on which the agent draws, in order to fashion an integrative act description, is material that would ordinarily be called his reasons for acting – the circumstances, motives, and other considerations that make one action rather than another the sensible thing to do'.[41]

In this two-stage process, not only do agents generate future actions on the basis of their self-conceptions and what would make sense to do in the light of them, but when they actually carry those actions out, they contribute to consolidating and bringing together those self-conceptions and their behaviour. In other words, we write a narrative for ourselves, but then we make that narrative true, in virtue of having it as a self-narrative. 'We invent ourselves … but we really are the characters whom we invent'.[42] So, confabulation can contribute to a sense of agency and self-consistency in the long term, which makes the ill-grounded self-ascription in the instance of confabulation truer over time.

Clearly, living up to positive self-ascriptions will be particularly valuable. Self-consistency motivations relate solely to being understandable and predictable. In confabulation, individuals have a further motivation to specifically protect and maintain *positive* self-concepts.[43] There will be further benefits when an agent's identity encompasses morally good self-representations such as being compassionate or generous, and these positive concepts can feature in confabulations. With these morally good self-representations protected and emphasised by confabulation, this can then trigger self-consistency motivations to drive the individual to live up to that conception in their behaviour and attitudes, as is described by Jefferson and Summers.[44] So, in protecting positive self-representations, confabulation protects a positive self-image to begin with, and although that might not be particularly accurate or warranted at that time, it can mean that individuals come to embody it by choosing actions in the future which cohere with this narrative.

The final benefits to discuss are those which follow from having shared an explanation at all, which others can scrutinise and comment on, despite it being a distorted one. Confabulation facilitates social interaction and maintenance of the social self.[45] More specifically, feedback received from others on the offered explanations can point out that they are getting something wrong – that they are ill-grounded. This is an epistemic gain for the individual and gets around what is for them a cognitive gap. For example, someone else may point out to the employer that the CVs are identical. This may be because others do not have the relevant self-serving motivations to protect the positive self-concepts of others (see Malle *et al*. for discussion).[46] So, despite not having introspective insight into the influence of some causal factor on their behaviour, they can come to learn about these indirectly, through confabulating, and then having them pointed out to them by others.

Relatedly, in responding to another's request for an explanation of behaviour and even beginning to formulate an explanation to share in the social world, considerations which were not in any way articulated before can be brought 'to the surface'. Considering things from the point of view of others, which is triggered once individuals bring their thoughts into the public realm, can begin this process. Not only do individuals find themselves thinking more explicitly about the story of their behaviour and its accuracy, but about the self-concepts they are drawing on within that story. As described above, this can then trigger self-consistency motivations to behave in keeping with that self-concept which has now been emphasised in the social world.

Despite the psychological and epistemic benefits described, one might think that much greater benefits would still be gained from simply having a more accurate appraisal of one's actions and the motivations behind them. But remember that giving an accurate explanation is not an option available to the agent who confabulates. The only other option is to be dumbfounded.

Now, one might think that the same (or greater) benefits could follow from being dumbfounded rather than confabulating. However, I think this is unlikely. If an agent

lacks the relevant motivations to confabulate (of seeking meaningful understanding, of social engagement, of identifying and communicating values, etc.) and to avoid being dumbfounded, then they are unlikely to have the motivation to reflect further on their lack of an answer. I see these epistemically valuable motivations more likely being employed at the first opportunity, in confabulating, rather than 'holding back' and kicking in later after being – quite unusually – comfortable to be dumbfounded at first, and in the presence of others. I have already established that there is a cognitive gap, and the dumbfounded agent is not sharing an explanation which could then be corrected by others in a way which can fill that gap for them. So even the thoughtful dumbfounded agent cannot come to realise that actually it was the stockings being on the right-hand side that made the difference. Granting as much as I can to this case, why and how could the dumbfounded agent suddenly start engaging more deeply with the question of the reasons for their behaviour? I can only imagine that they suddenly start to care about the topic and what it says about them, and/or they are persistently prodded by others. But these just are the triggers for confabulation materialising. The motivations are now in play; they want to protect self-representations and engage with the social world by giving an explanation. The dumbfounded person still has the cognitive gap and so still now has to go 'the long way round' – they have to piece together possible explanations, and then run through how likely they are. This is basically confabulation, but perhaps with more awareness of the process than usual, which may be stemming from the extra duress from others. It is natural to suppose that ordinarily, with the slightest interest in the matter at hand, this kind of process is what more swiftly gets going at the first opportunity.

More straightforwardly, agents admitting their ignorance on an important topic is frankly very rare. In many cases of confabulation, individuals are asked to explain a simple choice or behaviour, and these matters of basic self-insight are actually very important to individuals. In other cases, individuals can be asked about highly emotive topics such as ethical issues and political opinions, important topics which often reflect treasured identities and values (for discussion of confabulation in this latter context, see Murphy-Hollies).[47] Being short of answers here (for why one votes in a certain way, for example) is embarrassing, and an individual being comfortable with a lack of answer here would be – as described above – quite strange. This would, I suspect, have significant social repercussions and penalties. The person will be found strange and unpredictable. This social alienation is likely to bring exactly the kind of epistemic costs that confabulation has been described as securing; signalling that we are rational, cognisant actors who understand the reasons for our behaviour.

So, despite having limited and incomplete insight into the causes of one's own actions, being moved by the motivations which underlie confabulation, and therefore confabulating, can leave the agent better off than being dumbfounded. The forward-looking role of self-ascriptions discussed so far highlight how confabulation is not necessarily the end of enquiry, as the agent can – when a longer time-frame is considered – gain more accurate appraisals from this prompting of reflection and receipt of feedback from others. If individuals instead said nothing and were simply dumbfounded, they are unlikely to reflect on their behaviour and whether it coheres with some of their self-ascriptions. In this way, confabulation can open up, rather than close off, evaluations of behaviour which highlight unwarranted self-ascriptions and their poor match with the behaviour in question.

4. Applying These Benefits to the Development of Virtue

Individuals' ability to make changes to themselves (i.e. to align behaviour with certain traits and ascriptions) is clearly of importance to the development of virtue. Given that virtue is conceived of as an inherently developmental process[48] which is always maturing further as we learn from new people and situations, this will be of importance to everyone, from young, budding virtuous agents to wiser, more accomplished ones. In this section, I describe a few ways that these discussed benefits of confabulation can aid in the process of virtue development.

Velleman's account of acting for reasons describes individuals as seeking the appropriate *behaviours* for self-ascriptions, and this lays out how virtuous agents can make good intentions result in good behaviour. I have described how self-consistency motivations push individuals to live up to positive self-concepts, and once confabulations emphasise these self-concepts and make them more explicit to the agent, their influence on guiding future behaviour can be even stronger. For an ethical theory which centres traits as the most relevant ethical consideration, this is certainly a more promising way forward than confabulations just twisting all past circumstances into supporting positive self-concepts, regardless of how inaccurate that account of the circumstances is. This behaviour change also satisfactorily stems from *character*, as is important for virtuous behaviour; it must stem from a reliable disposition of character to act for certain sorts of reasons.[49] These two things, acting from character and for certain reasons, are linked closely in Velleman's account of acting for reasons.

Having an ongoing motivation to behave in ways which support self-concepts can predispose agents to notice when their present situation calls for the exercise of that virtue. For example, if someone quite consciously conceives of themselves as being a very compassionate person, this positive self-representation might be enhanced in their confabulations. With such an explicit and valued commitment to this self-concept, it seems harder for this person to do nothing when they hear about the suffering of some neighbours. They are at least more likely to notice that it is somewhat incongruous for them to be a compassionate person and yet not act in this circumstance. So, individuals are more likely to notice a requirement for virtuous behaviour, and to demonstrate it. By successfully engaging with a requirement for virtue in more of these scenarios, then over time agents accrue more and more practice of that virtue (of 'compassionate behaviour', for example) in the varied situations it can be manifested in. This contributes to making a habit of and cultivating that virtue, so that agents become competent and wise in demonstrating that virtue in all its nuances and varied applications.

So, I suggest that confabulation can help consolidate an agent's internal locus of control over their behaviour. Confabulation props up a positive self-concept even when it was not perfectly demonstrated, but this allows for it to continue to guide future behaviour. Self-consistency motivations work to align behaviour with self-ascriptions of virtue. When focusing on individual instances of confabulation, we see agents distort accounts of the past in order to prioritise the self-ascription over the reality. But looking forward, these self-ascriptions and the protection of them in turn protects their valuable role in determining future behaviour.

In the next section, I start to moderate this optimistic account of the role confabulation can play in the development of virtue, by considering the skills required in this complex process of aligning behaviour with ongoing self-narratives.

4.1. Self-Regulation: The Virtue

These benefits of confabulation are far from guaranteed or easy to acquire. We can imagine particularly stubborn and arrogant individuals who will not consider for a moment that their behaviour and attitudes do not reflect their positive self-representations, which are practically grandiose. These individuals just 'dig their heels in' in response to any other viewpoints, and any hope of gaining a better alignment between their self-ascriptions and behaviour is slim. In short, they persistently portray their actions in a biased way so as to preserve their grandiose self-concepts and do not do anything constructive with feedback they receive from others. I have described how having rose-tinted conceptions of oneself can trigger motivations to live up to it, but what if behaviour just does not change over time and reality is distorted again and again instead? In this case, self-consistency motivations seem to just drive further confabulations.

A systematic failure to align self-ascriptions and behaviour would be a critical obstacle to achieving deeply rooted, mature, and consistent virtuous behaviour. The effects of confabulation on an individual's ability to align self-ascriptions and behaviour are addressed by de Bruin and Strijbos.[50] They argue that confabulation does not necessarily inhibit this process (thereby allowing that it *can*), and they do so by appealing to self-ascriptions' forward-looking function. Like Velleman and Bortolotti, they also suggest that self-ascriptions allow agents to understand and predict themselves and their behaviour, thus making it more likely that individuals bring their behaviour into line with them over time.[51] They describe that the real concern for agents here is not single instances of failing to accurately capture factors which drove behaviour, as we see in cases of confabulation, but rather cases where individuals completely lose their status overall as reliable self-ascribers. This happens when someone consistently posits very poorly supported self-ascriptions over time, and shows little regard or interest in improving their skill and reliability at this. This is the situation which stubborn and arrogant individuals are in.

De Bruin and Strijbos outline a concept of 'self-know-how', which is a set of 'self-regulatory skills' and 'attitudes'.[52] These attitudes address specifically the practice of bringing behaviour more successfully into line with self-representations or, alternatively, updating self-ascriptions in the light of behaviour. Without these attitudes, agents may have an idea of what they are like which is completely disconnected from the reality, with no hope of gaining insight and re-alignment due to this systematic failure stemming from a lack of 'self-know-how'.

These attitudes include agents being flexible and curious with regard to the causes of their behaviour, being open-minded and receptive towards the perspectives of others, and being attentive to their own thoughts and feelings.[53] Agents with self-know-how have reasonable confidence in their self-ascriptions; not so little that they are plagued with doubt about their own self-evaluations and will not critically scrutinise how others describe them, but not so much that they are arrogant and completely dismiss the viewpoints of others.

I propose that the skills and attitudes encompassed in this notion of self-know-how make up a virtue of self-regulation. It is a sort of structural virtue which results in good overall management of the self and one's motivations,[54] such that other virtues can be realised and embodied. This is akin to descriptions of wisdom or phronesis as a meta-virtue,[55] because it works over and above any specific motivation or virtue, instead

managing and evaluating them from a meta-cognitive perspective. Another example of a structural virtue could be fortitude, for it is a trait which could be employed in the exercise of a number of different, more specific, motivations and virtues. Imagine how fortitude can help an individual be a constant source of support for a friend who is going through a very difficult time, who may not always be very receptive to that help despite being deserving of it, and no doubt very grateful one day. This will involve juggling many concerns, delicate judgements, difficulties, and setbacks, so that she can continue to be the best friend she can until her friend is in a better place. The same structural virtue, fortitude, may be employed in one's political activism, promise-keeping, or resilience to temptation. Structural virtues enable agents to structure and organise their lives with excellence; in a way which bolsters other virtues, such as kindness. In other words, they are self-governing traits which enable agents to achieve their aims.

When agents encounter some inconsistency in their understanding of themselves and the reality of their behaviour, they are called upon to recognise the need for earnest self-regulation and practice this virtue. This acts as a corrective, in Foot's sense, to the temptation of clinging onto one's treasured self-concepts in the face of opposing evidence, simply because the individual wishes to be that kind of person.[56] The virtue of self-regulation will also be ideal at the 'golden mean'. With too little self-regulation, individuals will have no open-mindedness and no curiosity, and will be completely dismissive of what others say to them, with their own word being the final word and the end of enquiry. With too much, individuals will be too preoccupied with the alignment of their self-ascriptions and behaviour such that they never try anything slightly new or 'out of character' for them. They would stringently avoid any self-ascriptions which could possibly be a little wishful or incorporate aspects of their idealised selves. This might mean that they confabulate less, but hopefully it is becoming clear what they can lose out on in doing this: the opportunity to enhance their positive self-representations such that they have a more powerful sway over future behaviour.

A final point is that as with other virtues, the virtue of self-regulation will not necessarily always lead the agent to the right action in all circumstances. For example, in the same way that courage can be mis-applied by the terrorist, self-regulation could technically be used to align behaviour better with *bad* self-concepts. Straightforwardly bad self-concepts such as being cruel and stingy will, however, not be in play in the context of exercising self-regulation in the aftermath of confabulation because confabulations are driven by motivations to preserve *good* self-concepts. Of course, this is 'good' by the lights of the agent, and this will very likely also depend on the social context which the agent is in and the traits which are 'good' in those contexts. So, this still leaves the issue of bad traits which are nevertheless valued or seen as good by the possessor, perhaps such as being 'macho' or 'domineering'. It is easy to see how certain social contexts will paint these traits as positive, and reward them. To give a full list of ways in which this could happen would be akin to listing all the ways in which something like 'courage' can go awry and be misapplied, and would be difficult. (Why and how can the courageous come to practise courage in misguided contexts?) But I think one of the most obvious and common ways that self-regulation could be mis-applied would be from the agent being situated within an unideal culture which encourages and values misguided ideals, and I tackle this issue in the final section of this article. In essence, though, this issue does not take away from the potential of confabulation, with self-regulation, to improve one's self-understanding, behaviour, and in turn, development of virtue.

Imagining the most egregiously self-deceiving individual you can really tests my picture. (Instances of liars and intentional deceivers are not the focus of this article as they are not confabulating.) It is difficult to see how them confabulating and casting their poor behaviour in a good light again and again could ever be a good thing. But I encourage us to think, even in these strongest cases, that these individuals go wrong not by confabulating but by not having the virtue of self-regulation. They confabulate constantly because they lack self-regulation, but at least in confabulating they continue to superficially hold themselves to good standards or values, and this is something that we as critics can grasp and hold them responsible for. They are opportunities which they keep missing, but opportunities none the less. In essence, confabulation can protect and prolong bad behaviour, but given the cognitive gaps in play, it also remains the only route to improvement.

4.2. Self-Regulation: The Skill

What is particularly pertinent to this virtue of self-regulation is the *know-how* which it draws upon. Just as de Bruin and Strijbos emphasise in their notion of self-know-how, these practices employed in re-aligning self-ascriptions and behaviour are skilful and rely not on propositional knowledge regarding the self, such as whether one is patient, rude, or impulsive, and so on. But rather it relies on practical knowledge of *how* to navigate one's own ascriptions, feelings, behaviours, and the viewpoints of others. The skill in self-know-how is in agents managing their own influence on their own evaluations of their behaviour. Instead of thinking that individuals simply 'read off', in a distanced and sterile manner, what their own traits are from the data that is their behaviour, it is better to recognise that as the authors of their own narratives and ascriptions they have influence and power to shape them in certain ways.

Accepting that individuals play a role in moulding what their self-ascriptions are by the very act of seeking to identify them and authoring them, reflects a move in the philosophical literature towards a mind*shaping* account of self-knowledge (and knowledge of others) from a mind*reading* account.[57] A mindreading account does not recognise this influence from the agent as the author of her own ascriptions; it is a more neutral process of accurately (or inaccurately) detecting actual ascriptions and mental states which are assumed to be there already, and therefore not in some sense partly formed or put there by the agent in the very act of looking inwards to understand oneself. De Bruin and Strijbos describe it thus:

'According to this [mindshaping] view, we are not primarily in the business of passively reading the mental states of others in order to predict or explain their behaviour. Rather, we are being socialised in a community held together by social pressures to make behaviour understandable. That is, we modify our own minds and those of others in accordance with the norms and normative expectations embodied in our community. This is effectuated by means of a variety of practices, behaviours and mechanisms – including imitation, pedagogy, norm cognition and enforcement, and language based regulative frameworks, like self- and group-constituting narratives'.[58]

So, in a way akin to personal self-narratives but on a wider societal scale, mental-state attributions on a mindshaping model have both a backward-looking function of making past behaviour understandable, and a forward-looking function of teaching people how to express themselves according to the norms of their folk psychology, and so shape future behaviour to be in keeping with that ascription and norms.

McGeer is an influential proponent of mindshaping in the context of shaping moral behaviour, and these practices of self-regulation in the aftermath of reason-giving provide support for her account of how agency and 'intelligent capacities' are formed in individuals.[59] McGeer's focus is on identifying when we can justifiably assign moral responsibility to an individual, arguing that the practice of having reactive attitudes towards others contributes to the formation of the very things which do make attributions of moral responsibility legitimate for that individual: the relevant agential and intelligent capacities. A particularly important and relevant intelligent capacity is being able to reliably respond to the right moral reasons, and McGeer describes this capacity as essentially developmental, with the degree to which an agent has mastered it varying over time.[60] It is also an extremely dynamic process which requires interaction with the environment and social feedback from other people in order to develop it in the first place and to sustain it over time.[61] I see the skill of self-regulation working in the same way.

She writes that 'the capacity to recognise and respond to moral reasons is an essentially social skill, requiring social feedback to develop and maintain',[62] and later that 'form[s] of moral address … are not in their nature merely backward-looking "reactive" attitudes; they are also forward-looking "evocative" attitudes, as we might say … calling for a particular response from the putative wrongdoer: for instance, to explain and/or justify what they have done; and where that fails, to acknowledge how they have failed to track and respond to the reasons that ought to govern their behaviour, and to commit to doing better in the future'.[63]

What is particularly relevant here to confabulation is that people are called upon to explain and/or justify their behaviour, acknowledge when it was not a response to the reasons which they are taking it to have been, and respond in such a way that their behaviour *will be* a response to those reasons and causes in the future. This capacity calls, first of all, for an explanation to be given by the agent for their behaviour. It does not have to be an accurate one. And so, confabulation can be valuable here and fill this role.[64] Second, this capacity comes in degrees and is essentially practice-dependent[65] and so calls for know-how. We can have propositional knowledge of the norms and values in our cultural milieu, but regulating ourselves and understanding others in the light of them requires something different from propositional knowledge: practical know-how. This is akin to the difference between knowing all the moves in chess, and being able to put them to use in playing a game. In terms of virtue, individuals can know propositionally *that* they ought to be kind and how kindness is often demonstrated and recognised, but it takes a know-how of self-regulation to be able to reliably and wisely enact being kind in the real world, such that their behaviour lives up to that self-ascription.

Given how common it is for individuals to confabulate, having the virtue of self-regulation will therefore be a very powerful factor in improving behaviour and realising virtue. More specifically, it is critical in determining whether confabulation brings benefits which can be applied to developing virtue, or whether confabulation is a missed opportunity. Or, at worst, a sustained lack of self-regulation can mean that some instance of confabulation is the final straw for an individual losing their overall status as a reliable self-ascriber.

In providing a useful and explicit skillset for aligning behaviour with self-concepts, self-regulation enables self-consistency motivations to be more effective. In the context of developing virtue, this skillset enables individuals to embody virtue more effectively over time. It keeps them sensitive to factors which are important for developing virtue: the

possibility of being wrong about themselves; that overlooked factors can influence their behaviour; that embodying positive traits is an ongoing process; and finally how best to navigate the input from others.

So, my suggestion is that budding virtuous agents should prioritise the cultivation of the virtue of self-regulation over attempting to somehow never confabulate. The development of virtue is already often compared with the cultivation of skill,[66] and opportunities to exercise and improve self-regulation present whenever individuals engage in reason-giving for past behaviour, even with imperfect explanations. Self-regulation and having this kind of constructive interest in achieving a high consistency between ascriptions and behaviour are what enable confabulation to positively enhance agency and the realisation of other virtues. Without the exercise of this virtue, which in turn aids in the development of other ones, individual instances of confabulation are missed opportunities for the psychological, epistemic, and moral benefits discussed. If this continues over time, agents lose their status as reliable self-ascribers and confabulation could have the undesired effect of assuring agents of their positive self-ascriptions despite the evidence against them. On the other hand, with self-regulation, agents are equipped with an ability to manage these inconsistencies in ways other than just rationalising them away in confabulations.

To summarise, I argue that being dumbfounded reflects a strange lack of interest and engagement from agents, and they are unlikely to gain benefits from this. If the agent confabulates and they do not have the virtue of self-regulation, they may dig their heels in and continue to believe a poorly supported self-ascription. If they do this for long enough, the worst-case scenario is that they lose their status as a reliable self-ascriber. If they confabulate and do have the virtue of self-regulation, then through further reflection and prompts from others, they can realise the unendorsed influence of extraneous factors on their behaviour and rectify it over time. We see that their behaviour lives up to the protected self-ascription in the confabulation, making it true. Where agents go wrong is not in confabulating, but in not exercising the virtue of self-regulation afterwards.

5. A Challenge

Adopting a mindshaping account of understanding oneself and others, as opposed to a mindreading account, allows us to see that it is hasty to write people off as having very poor self-knowledge and being incapable of deeply rooted virtuous behaviour, because of confabulation and the poorly supported reasons people often give for their behaviour. It is to overlook and dismiss their ability to go on to practise self-regulation and address this discrepancy. As we have seen, confabulation can be a way that attention is drawn to overlooked causal determinants of behaviour and managed in a way which does not depend on solitary, explicit deliberation about oneself which can so often be inadequate.[67] Instead, it draws on the valuable input of others to fill in our 'blind-spots'. A mindshaping account of ascriptions emphasises their *forward-looking* function in regulating behaviour, so singular instances of poor self-ascriptions are not the end of the road.

However, embedding reason-giving practices so fundamentally within shared folk psychological norms can bring a problem. I have described how confabulations can prompt

feedback from others which helps agents better understand their behaviour. However, all of an agent's peers are likely to be in the same folk psychological and cultural milieu as themselves. If agents are providing representative reasons for their behaviour which they have learnt are acceptable by the standards of their current folk psychology, it is easy to see how this could provide questionable standards for behaviour. This is a key way in which agents might end up seeing some of their disruptive and unpleasant traits as positive. The norms and values in some folk psychological milieus can be misguided; patriarchal representations might be over-valued and behaviour which should be seen as violent and controlling is seen as 'protective' or embodying 'leadership'. Similarly, vulnerable groups may learn to receive and express love in ways which are in fact exploitative and undermining. Finally, political propaganda can quite successfully classify some norms and behaviours as embodying 'freedom' and 'justice' when this is not really the case. The possibility of being in a culture which is confused in this way seems to limit the value that confabulations and self-regulation can bring. How could budding virtuous agents combat this?

My response to this concern is two-fold. First, I suggest that self-regulation could still help agents see and realise that there are inconsistencies in values and behaviour, even at the societal or cultural level. Applying the skill of self-regulation can mean that members of oppressive or fascist political systems which are saturated with propaganda come to realise that their behaviours and cultural practices do not really embody the values they are espousing. After all, there are plenty of groups of cultural critics and protesters who argue that certain policies or valued customs do not actually make people free or empowered. And it is important to note that the actions of these people and groups can be very effective in bringing about widespread change and modification of norms. One reason for this is that self-regulation is not a one-way street.[68] In changing our own behaviour (or self-concepts) in order to maintain consistency and live up to those descriptions, we implicitly change standards for the people around us as well.[69] The way we interact with others will change, which, in turn, changes how those interact back with us, in a sort of social looping effect.[70] Within this dynamic of incorporating the new ways that others behave and interact with us into our latest self-concepts, self-regulation is still a crucial and helpful tool. Not only this, it can even help in regulation of values beyond the scope of just the individual, to chip away at problematic aspects that make up the unideal culture itself.

Second, the drawbacks of being situated in unideal cultures is not something which is completely absent in established ideas about how virtue develops. Virtue ethicists already emphasise the need for certain features to be present in the environment for individuals to have a good 'moral education' and for virtue to develop to full maturity. Annas describes how the virtues are not learnt primarily from abstract thought and reflection, but from 'embedded contexts'; the experiences individuals have within their environment.[71] They rely on opportunities being available for the natural 'roots of virtue' within them to develop and mature.[72] Furthermore, individuals greatly benefit from having moral role models to guide them.[73] So, there will be an element of luck in whether agents have access to these things, in the same way that there will be an element of luck in whether agents find themselves in cultures with particularly good norms and practices which really reflect and express certain values and ideals. In essence, I bite this bullet and accept that these agents are in trouble, but I also maintain that having the virtue of self-regulation gives them the best chance of getting out.

6. Summary and Conclusions

In this article I have discussed how some of the suggested pragmatic and epistemic benefits of confabulation[74] can be conducive to the development of virtue. However, effectively gaining these benefits and making the desired changes will require exercising the meta-virtue of self-regulation. This enables us to be effective and successful at making our behaviour consistent with virtuous positive self-representations. Confabulation is an opportunity to practise self-regulation and make the self-concepts and self-ascriptions emphasised in those unideal explanations truthful.

Accepting mindshaping and situating the practice of reason-giving so deeply in the context of social life and social interaction means that confabulation can be, over time, a social tool for addressing unwarranted self-ascriptions and overlooked causal influences on our behaviour. This is a conception of confabulation as a social and communicative practice which, with to-and-fro with others and over time, can produce more robust, consistent, and deeply understood virtuous behaviour. A conception such as this shows us how confabulation can be a beneficial and even expectable part of the process of becoming virtuous. To lose confabulation would be to lose a motivation and interest in gaining insight and understanding of ourselves and what drives our behaviour. Confabulation is part of the process of getting better at more accurate reason-giving, and particularly if an agent has the virtue of self-regulation, they can more effectively align their behaviour with virtuous self-concepts. Another effect of recognising confabulation as a natural social tool for getting better at identifying the reasons for our behaviour over time, and responding to the reasons we want to, is accepting this as a perfectly legitimate route to virtue. McGeer's account of mindshaping draws our attention to how much of a dynamic negotiation reason-having and reason-giving is, and that previous conceptions of ideal virtue development, which draw on having perfectly well-grounded and pre-established reasons for behaviour, may be unrealistic.

Kathleen Murphy-Hollies, University of Birmingham, Birmingham, UK. *k.l.murphy-hollies@bham.ac.uk*

Acknowledgements

I owe a lot of people a lot of gratitude for help with developing the ideas in this article. Heartfelt thanks go to Lisa Bortolotti, Iain Law, and Quassim Cassam. Also to Fabienne Peters for extensive discussion on an earlier draft, and Tom Baker for very helpful comments. Thanks also to audiences at the Women in Philosophy Networking and Mentoring event at the University of Birmingham (2022), the Philosophy at the Intersection of Moral Responsibility, Agency and Regulation postgraduate conference at the University of Birmingham (2022), the Society for Applied Philosophy Annual Conference in Edinburgh (2022), the Reasons, Rationality and Culture Workshop at Tilburg University (2021), and the Understanding Value X conference at Sheffield University (2021). And finally, thank you to the philosophy department at the University of Birmingham for discussing these ideas with me at departmental talks and at other times, for fun.

NOTES

1 Hirstein, *Brain Fiction*, 204; Bortolotti and Cox, "'Faultless' Ignorance," 957.
2 Bortolotti, "Stranger than Fiction."
3 McGeer, "Mind-Making Practices"; McGeer, "Scaffolding Agency."
4 Scaife, "Problem for Self-Knowledge"; Carruthers, *Opacity of Mind*.
5 De Bruin and Strijbos, "Confabulation"; Andreotta, "Confabulation."
6 Nisbett and Wilson, "Telling"; Hall *et al.*, "Magic at the Marketplace."
7 Haidt, "Emotional Dog."
8 Wilson *et al.*, "Introspecting."
9 Bortolotti and Cox, "'Faultless' Ignorance," 957; Hirstein, *Brain Fiction*, 205.
10 Sullivan-Bissett, "Implicit Bias," 551.
11 Nisbett and Wilson, "Telling."
12 Ibid., 243–4.
13 Bergamaschi Ganapini, "Confabulating Reasons"; Sullivan-Bissett, "Implicit Bias," 552.
14 Bertrand and Mullainathan, "Emily and Greg."
15 For a discussion of the distinction between motivating and causal reasons, and its implications for confabulation research, see Sandis, "Verbal Reports." Sandis argues that participants often, reasonably, assume that investigators are asking them for their motivating reasons for their choices and behaviours, and therefore participants are sensible to provide these and to overlook causal influences which, even if efficacious, were not their reasons for their choice/behaviour. I acknowledge that this is an important point, but I hope to emphasise in this article the repercussions of motivating reasons which make false claims about the world, and of failures to identify some important causal reasons too.
16 Sullivan-Bissett, "Implicit Bias," 552.
17 Holroyd, "Implicit Bias, Awareness and Imperfect Cognitions."
18 Holroyd, "Responsibility," 275.
19 Nisbett and Wilson, "Telling."
20 For most cognitive gaps in everyday situations, whether it be in an experiment or not, the gaps will stem from decisions having been the output of emotional system 1 processing, and this type of cognitive processing is simply not introspectively accessible (Evans, "Dual-Processing Accounts"). However, some experiments do appear to engineer particularly large cognitive 'gaps', as in the case of choice blindness studies. In these cases, participants not only fill the gap of why they made a certain choice, but also what the choice itself actually was, because they give reasons for a choice which they did not make (following manipulations of the experiment). I am grateful to an anonymous reviewer for pointing out this possible variation in how extreme these gaps might be.
21 Haidt, "Emotional Dog."
22 Dent, *Moral Psychology*, 7.
23 Aristotle, *Nicomachean Ethics*, II.6.
24 Ibid., VI.13, 1144b.
25 Hursthouse, *On Virtue Ethics*, 124.
26 Ibid., 11–12.
27 Williams, "Acting"; Audi, "Acting."
28 Hursthouse, *On Virtue Ethics*, 129.
29 Annas, *Intelligent Virtue*, 19.
30 Swanton, "Virtue Ethical Account."
31 Sullivan-Bissett, "Implicit Bias," 552.
32 Ibid., 552.
33 Bergamaschi Ganapini, "Confabulating Reasons."
34 Örulv and Hydén, "Confabulation."
35 Stammers, "Confabulation."
36 Coltheart, "Confabulation."
37 Bortolotti, "Stranger than Fiction," 241.
38 Ibid., 242.
39 Ibid., 242.
40 Velleman, "From Self Psychology."
41 Ibid., 367.

42 Velleman, "Self as Narrator," 206.
43 Sullivan-Bissett, "Implicit Bias," 552.
44 Jefferson, "Confabulation"; Summers, "Post Hoc."
45 Bortolotti, "Stranger than Fiction," 243; Sullivan-Bissett, "Implicit Bias," 555.
46 Malle *et al.*, "Actor–Observer Asymmetries."
47 Murphy-Hollies, "Self-Regulation."
48 Annas, *Intelligent Virtue*, 38.
49 Hursthouse, *On Virtue Ethics*, 11–12.
50 De Bruin and Strijbos, "Confabulation."
51 Ibid., 155.
52 Ibid., 158–9.
53 Ibid., 158–9.
54 Adams, *Theory of Virtue*.
55 Kristjánsson *et al.*, "Phronesis."
56 Foot, *Virtues and Vices*.
57 See Zawidzki, "Function of Folk Psychology"; Zawidzki, *Mindshaping*; De Bruin, "First-Person Folk Psychology"; McGeer, "Moral Development"; McGeer, "Mind-Making Practices."
58 De Bruin and Strijbos, "Confabulation," 153.
59 McGeer, "Scaffolding Agency."
60 Ibid., 309, 311.
61 Ibid., 311–2.
62 Ibid., 313.
63 Ibid., 314.
64 Similarly, I suggest that reasons given in confabulation can play the same kind of role that Hutto (*Folk Psychological Narratives*) describes in his Narrative Practice Hypothesis: that it is through the sharing of one's reasons for action, and the narratives those actions are placed into, that children learn the basics of acting for a reason and attributing simple mental states to others.
65 McGeer, "Mind-Making Practices," 262.
66 Annas, *Intelligent Virtue*, 16.
67 Strijbos and de Bruin, "Self-Interpretation," 304.
68 Many thanks to an anonymous reviewer for pointing out this important extra mechanism in the dynamic between changing our own behaviour and the behaviour of those around us.
69 De Bruin, "First-Person Folk Psychology," 181.
70 Andrews, "Pluralistic Folk Psychology," 284.
71 Annas, *Intelligent Virtue*, 21.
72 Ibid., 31.
73 Hursthouse, *On Virtue Ethics*, 35.
74 Bortolotti, "Stranger than Fiction."

References

Adams, Robert Merrihew. *A Theory of Virtue: Excellence in Being for the Good*. Oxford: Clarendon Press, 2008.
Andreotta, Adam J. "Confabulation Does Not Undermine Introspection for Propositional Attitudes." *Synthese* 198, no. 5 (2021): 4851–72.
Andrews, Kristin. "Pluralistic Folk Psychology and Varieties of Self-Knowledge: An Exploration." *Philosophical Explorations* 18, no. 2 (2015): 282–296.
Annas, Julia. *Intelligent Virtue*. Oxford: Oxford University Press, 2011.
Aristotle. *The Nicomachean Ethics*. Oxford: Oxford University Press, 2009.
Audi, Robert. "Acting From Virtue." *Mind* 104, no. 415 (1995): 449–471.
Bergamaschi Ganapini, Marianna. "Confabulating Reasons." *Topoi* 39, no. 1 (2020): 189–201.
Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94, no. 4 (2004): 991–1013.

Bortolotti, Lisa. "Stranger than Fiction: Costs and Benefits of Everyday Confabulation." *Review of Philosophy and Psychology* 9, no. 2 (2018): 227–249.
Bortolotti, Lisa, and Rochelle E. Cox. "'Faultless' Ignorance: Strengths and Limitations of Epistemic Definitions of Confabulation." *Consciousness and Cognition* 18, no. 4 (2009): 952–965.
Carruthers, Peter. *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. New York: Oxford University Press, 2011.
Coltheart, Max. "Confabulation and Conversation." *Cortex* 87 (2017): 62–68.
de Bruin, Leon. "First-Person Folk Psychology: Mindreading and Mindshaping." *Studia Philosophica Estonica* 9 (2017): 170–183.
de Bruin, Leon, and Derek Strijbos. "Does Confabulation Pose a Threat to First-Person Authority? Mindshaping, Self-Regulation and the Importance of Self-Know-How." *Topoi* 39 (2020): 151–161.
Dent, Nicholas John Henry. *The Moral Psychology of the Virtues*. Cambridge, UK: Cambridge University Press, 1984.
Evans, Jonathan St B. T. "Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition." *Annual Review of Psychology* 59 (2008): 255–278.
Foot, Philipa. *Virtues and Vices*. New York: Oxford University Press, 2002.
Haidt, Jonathan. "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108, no. 4 (2001): 814–834.
Hall, Lars, Petter Johansson, Betty Tärning, Sverker Sikström, and Thérèse Deutgen. "Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea." *Cognition* 117, no. 1 (2010): 54–61.
Hirstein, William. *Brain Fiction: Self-Deception and the Riddle of Confabulation*. Cambridge, MA: MIT Press, 2005.
Holroyd, Jules. "Implicit Bias, Awareness and Imperfect Cognitions." *Consciousness and Cognition* 33 (2015): 511–523.
Holroyd, Jules. "Responsibility for Implicit Bias." *Journal of Social Philosophy* 43, no. 3 (2012): 274–306.
Hursthouse, Rosalind. *On Virtue Ethics*. Oxford: Oxford University Press, 1999.
Hutto, Daniel D. *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge: MIT Press, 2012.
Jefferson, Anneli. "Confabulation, Rationalisation and Morality." *Topoi* 39, no. 1 (2020): 219–227.
Kristjánsson, Kristján, Blaine Fowers, Catherine Darnell, and David Pollard. "Phronesis (Practical Wisdom) as a Type of Contextual Integrative Thinking." *Review of General Psychology* 25, no. 3 (2021): 239–257.
Malle, Bertram F., Joshua M. Knobe, and Sarah E. Nelson. "Actor–Observer Asymmetries in Explanations of Behavior: New Answers to an Old Question." *Journal of Personality and Social Psychology* 93, no. 4 (2007): 491–514.
McGeer, Victoria. "Mind-Making Practices: The Social Infrastructure of Self-Knowing Agency and Responsibility." *Philosophical Explorations* 18, no. 2 (2015): 259–281.
McGeer, Victoria. "The Moral Development of First-Person Authority." *European Journal of Philosophy* 16, no. 1 (2008): 81–108.
McGeer, Victoria. "Scaffolding Agency: A Proleptic Account of the Reactive Attitudes." *European Journal of Philosophy* 27, no. 2 (2019): 301–323.
Murphy-Hollies, Kathleen. "Self-Regulation and Political Confabulation." *Royal Institute of Philosophy Supplements* 92 (2022): 111–128.
Nisbett, Richard E., and Timothy D. Wilson. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84, no. 3 (1977): 231–259.
Örulv, Linda, and Lars-Christer Hydén. "Confabulation: Sense-Making, Self-Making and World-Making in Dementia." *Discourse Studies* 8, no. 5 (2006): 647–673.

Sandis, Constantine. "Verbal Reports and 'Real' Reasons: Confabulation and Conflation." *Ethical Theory and Moral Practice* 18 (2015): 267–280.
Scaife, Robin. "A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously." *Acta Analytica* 29 (2014): 469–485.
Stammers, Sophie. "Confabulation, Explanation, and the Pursuit of Resonant Meaning." *Topoi* 39, no. 1 (2020): 177–187.
Strijbos, Derek, and Leon de Bruin. "Self-Interpretation as First-Person Mindshaping: Implications for Confabulation Research." *Ethical Theory and Moral Practice* 18 (2015): 297–307.
Sullivan-Bissett, Ema. "Implicit Bias, Confabulation, and Epistemic Innocence." *Consciousness and Cognition* 33 (2015): 548–560.
Summers, Jesse S. "Post Hoc Ergo Propter Hoc: Some Benefits of Rationalization." *Philosophical Explorations* 20, no. sup1 (2017): 21–36.
Swanton, Christine. "A Virtue Ethical Account of Right Action." *Ethics* 112, no. 1 (2001): 32–52.
Velleman, J. David. "From Self Psychology to Moral Philosophy." *Philosophical Perspectives* 14 (2000): 349–377.
Velleman, David. "The Self as Narrator." In *Self to Self: Selected Essays*, 203–223. Cambridge: Cambridge University Press, 2006.
Williams, Bernard. "Acting as the Virtuous Person Acts." In *Aristotle and Moral Realism*, edited by R. A. Heinaman, 13–33. Boulder: Westview Press, 1995.
Wilson, Timothy D., Douglas J. Lisle, Jonathan W. Schooler, Sara D. Hodges, Kristen J. Klaaren, and Suzanne J. LaFleur. "Introspecting About Reasons Can Reduce Post-Choice Satisfaction." *Personality and Social Psychology Bulletin* 19, no. 3 (1993): 331–39.
Zawidzki, Tadeusz W. "The Function of Folk Psychology: Mind Reading or Mind Shaping?" *Philosophical Explorations* 11, no. 3 (2008): 193–210.
Zawidzki, Tadeusz Wieslaw. *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press, 2013.

Paper 2: Stories as evidence

Murphy-Hollies, K. and Bortolotti, L. (2021), 'Stories as Evidence', in Memory, Mind & Media, 1(3). DOI: https://doi.org/10.1017/mem.2021.5.

Memory, Mind & Media (2022), **1**, e3, 1–12
doi:10.1017/mem.2021.5

MMM
Memory, Mind & Media

RESEARCH ARTICLE

Stories as evidence

Kathleen Murphy-Hollies and Lisa Bortolotti

Philosophy Department, University of Birmingham, Edgbaston B15 2TT, UK
Corresponding author: Lisa Bortolotti, email: l.bortolotti@bham.ac.uk

(Received 9 September 2021; accepted 14 September 2021)

Abstract

People often use personal stories to support and defend their views. But can a personal story be evidence? A story tells us that a certain event can happen and has already happened to someone, but it may not always help us understand what caused the event or predict how likely that event is to happen again in the future. Moreover, people confabulate. That is, when they tell stories about their past, they are likely to distort reality in some way. When people who lack access to what motivated past behaviour are asked why they made a choice, they tend to offer plausible considerations in support of that choice, even if those considerations could not have played a motivating role in bringing about their behaviour. When people experience impairments in autobiographical memory, they tend to fill the gaps in their own story by reconstructing significant events to match their interests, values, and conception of themselves. This means that people often offer a curated version of the events they describe. In this paper, we argue that the pervasiveness of confabulation does not rule out that personal stories can be used as evidence but invites us to reflect carefully about what they are evidence of. And this is especially important in the context of digital storytelling, because stories shared on online platforms can exert even greater influence on what people think and do.

Keywords: stories; evidence; memory; confabulation; social media; arguments; explanation; justification

Introduction

Memory is a powerful source of knowledge. When we tell stories about our past, we share significant experiences with others. Sometimes, we use personal stories to support a stance. A man who believes that the threat posed by the coronavirus pandemic has been widely exaggerated may say that COVID-19 is 'just like the flu' because when *he* tested positive for the virus, his symptoms were a runny nose and a bit of coughing. A general conclusion stems from a specific event.

In some contexts, sharing a story is an important move in a public debate. It may be used to discredit a political opponent or to enhance the credibility of a political ally. During a rally in April 2016, Donald Trump claimed that on 9/11, he helped responders at Ground Zero after the attack to the Twin Towers. With such a story, Trump presented himself as a man of action, generous with his time. However, there are no reports about Trump being at Ground Zero or helping responders on the day of the attack – and indeed

he himself said on a different occasion that he was at home when the planes hit the Towers (Voytko 2020).

Does it matter whether Trump's story is true? We tend to think that it does. If Trump fabricated the story to present himself as a trustworthy leader and a generous person, then the story could backfire, proving that he is happy to lie to enhance his image. But many people genuinely misremember the details of what they did during an unusually emotional or significant event, including 9/11. If Trump did not lie and simply misremembered what happened on that day, then the story would not tell us much about what kind of person Trump is but what kind of person *he thinks* he is or what kind of person *he wants to* be. He sees himself as someone who would help the responders after the attack.

The fact that people are vulnerable to memory distortions matters to the project of evaluating stories as evidence. But how? In the section 'Personal stories as evidence', we briefly consider some features of stories that affect their use as evidence. In the section 'Digital storytelling', we reflect on how social media change the way stories are shared and consumed. In the section 'Confabulation', we show how confabulation compromises the reliability of personal stories in clinical and non-clinical contexts. Finally, in the section 'What are stories evidence of?', we point to the positive and negative aspects of the use of personal stories as evidence.

Personal stories as evidence

In debates that have involved citizens on social media – such as what measures are most effective in tackling the coronavirus pandemic, whether the Earth's climate is changing due to human intervention, or whether electro-convulsive therapy is safe and effective – personal stories are often used as evidence for defending some of the debated claims. This invites us to consider some important features of stories.

Relevance

During the debate about Brexit in the UK, many people campaigning for the UK to leave the EU or justifying their decision to vote leave after the referendum used their own negative experiences with foreign immigrants to support their stance (Stammers and Bortolotti 2020). Experiences of delayed access to healthcare and fewer employment opportunities for British people were interpreted as an effect of excessive and unregulated immigration from the EU. In most cases, though, the stories shared to prove the point did not highlight problems specific to immigration from the EU and did not consider other factors that might have contributed to the problems, such as cuts to healthcare and other austerity measures.

Editing

Consider the debate about vaccine safety. Many people who oppose immunisations base their arguments on stories of children experiencing ill health, even death, after being vaccinated. Baby Ian Gromowski died after his hepatitis B immunisation. This case is often mentioned as evidence that parents need to consider the risks when agreeing to their children being immunised. The possibilities that other factors might have led to Ian's death are not usually part of how anti-vaccination groups report the story, such as the fact that the baby's mother had a difficult birth, that the baby was allergic to antibiotics, and that he had contracted a virus prior to receiving his immunisation (Shelby and Ernst 2013). Omitting such details may affect how the story is received and enhance its influence as evidence against vaccine safety.

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

Causal relationships

Audiences are not typically alert to the fact that personal stories are used as evidence. We all tend to focus on the relational and aesthetic features of a story rather than its accuracy. In other words, when we attend to a story, we may feel engaged, entertained, or moved if the story is gripping, funny, or touching. However, participants in a public debate need to do more than that: they need to be able to develop a critical attitude towards the story on the basis of whether it succeeds in supporting or challenging the debated claims. Are delays in healthcare provision due to the UK being in the EU? Was baby Gromowski's death caused by his hepatitis B immunisation? Unless answers to questions about causal relations between events can be answered, the story may fail to provide conclusive or even convincing evidence for the claim it allegedly supports.

Motivation

When we attend to personal stories, we are likely to be motivated by them, more so than by other forms of evidence. This gives rise to the *identifiable victim effect* (Jenni and Loewenstein 1997). In the context of charitable donations, people are more likely to act on the basis of the vivid recount of an individual's adverse experience than by statistical information informing them about a threat for a far greater number of people. This is because people prefer to act to stop a threat that is certain as opposed to a threat that is just probable, and they are sensitive to the percentage of individuals they can save, given the appropriate reference class. Donors reading the case of a young African girl dying of malnutrition believe that by donating to the cause they will *definitely* save *that girl*. Statistical information about a third of African children dying of malnutrition is not equally motivating because potential donors feel that the outcome is less certain.

Windows into another world

Sharing a personal story is mostly apt to capture a person's perspective on an event or an experience that may not be shared by others – the story may be the only source of information about what storytellers are going through when their situation is different from their audience's. Take first-person accounts of ill-heath or disability: such stories make it possible for everyone to understand the challenges caused by a debilitating condition and better inform support services (Bortolotti and Jefferson 2019). Another example is when people belonging to an ethnic minority report instances of racism in the place where they work: their stories are good evidence for the claim that there is a problem with racism that needs to be addressed in that workplace, a problem that others may not have been able to identify on their own.

Reliability

Whether the story is an accurate representation of reality depends on the storyteller's perceptual, introspective, and inferential capacities, and also on their standpoint and character. Did they directly experience the information they are sharing? Did they have special access to that information? Are they being intellectually honest? All of these questions are important, but it is difficult to address them in general terms. In the rest of this paper, we shall focus on a reliability question that applies to all cases in which people tell personal stories: is the storyteller vulnerable to *confabulation*? People confabulate when, without any intention to deceive, they describe or explain an event when they have no access to the evidence that would support the accurate description or explanation of that event (Bortolotti 2018). Confabulation is part and parcel of how people reconstruct their past and we will return to it in the section 'Confabulation'.

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

Digital storytelling

One important issue in considering stories as evidence is how the medium used to share the story affects its content, structure, and reach. Storytellers can make use of different tools seamlessly in digital storytelling, using text updates, hashtags, photos, links, GIFs, videos, sounds, and emoticons to share their experiences but also to participate in a debate or campaign for a cause. How do personal stories shared online differ from stories shared in the pub or at work?

Digital stories are co-constructed

Online stories are typically co-constructed. In something like a Facebook status update, people *co-construct storyworlds* because multiple storytellers are engaged via the practices of commenting, liking, linking, tagging, and sharing photos (Page et al 2013). This enables people who may be geographically or culturally apart to develop a shared story that contributes to their identities. The sharing of personal stories is identified as the main purpose of Twitter, but story ownership is complicated on social media, as the telling of a personal experience often involves multiple storytellers and the plotline is subject to changes during the telling and sharing of the story (Dayter 2015). While these social media practices undoubtedly bring people together, and encourage social interaction and group identity, determining authorship and authenticity may be more of a challenge for digital storytelling. A story can be told first by one user for one purpose and then hijacked for a different purpose by other users in a way that is not always transparent to the wider audience.

Digital stories are identity building

The same type of content, say *images*, can also be made to play different roles and tell distinct aspects of a story. The successful use of Instagram by a presidential candidate was analysed in the 2016 Austrian election campaign (Liebhart and Bernhardt 2017). The content posted on the platform had different purposes: document the campaign; demonstrate people's support for the candidate and the candidate's approachability; report the candidate's media work; call people to action, e.g. to make donations; provide a background story for the candidate, revealing him to be authentic as a person with certain interests and values; clarify the candidate's position on a number of important issues.

In a study of people posting images on Instagram and Snapchat during a visit to a museum (Villaespesa and Wowkowych 2020), there is also a wide range in visitors' main motivations for sharing; communicating the emotions they felt when seeing the artwork; getting other people to appreciate art and visit the museum; expressing their identity – which includes supporting something that is important to the person they feel they are. For instance, one visitor said:

> Then this was Rosa Parks, you know the bus situation, sitting in coloreds only, sitting in the back of the bus and things like that. And this particular photo I like, of course because I am African American and my father actually marched with Dr. Martin Luther King.

Digital stories have a wider reach

One observation made both by researchers interested in the political uses of social media and by those studying the effects of sharing aesthetic experiences is that digital storytelling sparks interactions faster and on a bigger scale. As we saw, the story can have different aims, is actively enriched by participants on the relevant social media platform, and

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

can be interpreted and shared further in ways that are not consistent with the original storyteller's intentions. Such features are not unique to digital storytelling. However, as misinformation experts have emphasised (e.g., O'Connor and Weatherall 2019), the influence of digital storytelling is far-reaching and more amplified than the influence of traditional storytelling, because the story is available, immediately, to a much bigger audience.

This explains the quick spreading of conspiracy theories in the age of social media: although there has always been a tendency for people towards conspiratorial ideation, especially during times of crisis, the impact of conspiracy theories shared globally via digital storytelling is greater. Thus, when the story is supposed to serve as evidence, it is important to examine the reliability of the source.

Confabulation

In the section 'Personal stories as evidence', we talked about how personal stories can fail at supporting a general claim. One of the reasons for concern was the tendency people have to confabulate, that is, to reconstruct their past in a way that others find inaccurate or badly supported by evidence. Here, we would like to say more about confabulation, starting with two examples:

> *Interview*: An employer is hiring for a new position and is looking through the CVs of the applicants. She chooses to invite George for an interview instead of Rabia. When the employer is asked by the panel why she chose George instead of Rabia, she says that George's CV was better than Rabia's. However, the candidates' CVs were of comparable strength. The employer does not have introspective access to any implicit biases that may have influenced her choice to interview George rather than Rabia and gives an answer by which she seemingly justifies her decision and her belief that George is a better choice.
> *Care Home*: A patient with dementia is in a care home. She explains her unfamiliar surroundings by saying: 'I am at work'. When she is asked by another resident to explain the presence of nurses and family members, she elaborates: 'Well, we are having renovations and building work done, these people are helping'. She does not have access to the relevant information that the building is a care home, that she has dementia, and that those people are there to keep her safe, but she has an explanation of the events that works for her and she sincerely endorses.

A common case of confabulation is when a person offers an explanation for why they acted in a certain way without being aware of the factors responsible for that action. Often, confabulations distort reality or misrepresent the specifics of the person's circumstances. Crucially, confabulators are not aware of the falsity or ill-groundedness of their explanations and therefore cannot be said to have any intention to deceive. They tend to be confident in their confabulations and are often, but not always, resistant to updating their explanations in the face of counter-evidence (Bortolotti and Cox 2009, 956).

In *Interview*, the employer displays non-clinical, everyday confabulation. Her confabulation is provoked because, without having been questioned, she may not have considered why she chose George over Rabia. Provoked confabulations tend to be momentary, whereas spontaneous confabulations are more likely to be persistent, as there may be a consistent 'outpouring' of false memories (Bortolotti and Cox 2009, 953). The latter is more commonly associated with confabulation seen in the context of psychiatric illness.

In *Care Home*, the patient shows spontaneous clinical confabulation because she offers the initial explanation of the unfamiliar surroundings unprompted and continues to

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

explain subsequent events with further confabulations. The confabulation may be a core symptom of the illness or a way to explain sensations and beliefs occurring to the patient because of their condition. Confabulation is commonly associated with Korsakoff's syndrome, amnesia, and anosognosia (Hirstein 2005, 8). In *Care Home* primary and secondary confabulations are both exhibited; in a manner akin to storytelling, the patient's initial confabulation that she is at work is developed with further claims about the presence of family and nurses, which would be explained by renovation works being carried out.

Neither of the confabulatory explanations is based on the features of the situation for which the explanation is offered. In *Interview*, the CVs of the job candidates were of comparable strength, and so the claim that George's CV was better than Rabia's cannot explain how the employer came to her decision. In *Care Home*, the unfamiliar surroundings do not belong to a workplace, and renovations are not in progress, and so these statements do not account for why the patient finds herself in a home surrounded by nurses and family members.

It is a curious but defining feature of confabulation that the person does not realise that they do not know what caused their experience or what has driven their behaviour. In *Interview*, the employer lacks access to the accurate explanation of her choice, because she does not have an insight into her implicit bias in favour of George. In *Care Home*, due to dementia, the patient lacks access to the autobiographical events that would support an accurate explanation of her situation – she does not realise that she needs to be cared for in a care home.

Why are people so keen to 'fill in' gaps in their understanding of reality and of themselves, despite not having access to the relevant evidence? A number of possible reasons are discussed in the literature. People may be motivated to preserve a sense of agency, a sense that they can influence what happens around them and act in such a way as to fulfil their goals. They have a desire not to be dumbfounded and humiliated by not being able to articulate the reasons for their own behaviour (Sullivan-Bissett 2015, 552). Moreover, people may fill the gaps in the way they do because they are also motivated by the desire to preserve positive self-representations. The employer in *Interview* wants to be a fair and competent decision maker who is neither sexist nor racist. The dementia patient in *Care Home* sees herself as a self-sufficient, independent person as opposed to a person facing debilitating illness and needing care (Bortolotti and Sullivan-Bissett 2018).

In some situations, the consumption of digital stories that resonate with one's values makes it more likely that there are confabulations readily available to use when an explanation is needed. Consider a third example, the case of some college students who are invited to explain why they do not wear face masks during the pandemic.

> *Mask Mandate*: A student is prevented to access college buildings because their institution has implemented a state-wide mask mandate. Together with some friends, the student starts a protest and is asked by the local press: 'Why are you protesting? Why don't you want to wear a mask?'. Students respond by availing themselves of explanations and justifications that have been offered on social media by political leaders and campaigners in the previous weeks. When making sense of their own choices, they express their views along these lines: 'It is our faces, we decide what to do with them', 'Ours is a fight for freedom', 'We do not belong to the government or to the college'. They tell a story where the refusal to wear a mask is a point of principle, namely a defence of their individual freedom. The factors leading to their behaviour may be different, though, and include considerations about convenience, a desire for non-conformity, denial of the health threat, scepticism about the effectiveness of masks, or a desire for normality, just to mention a few. (For the real-life case inspiring this example, see Smith 2021.)

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

In general, people have a strong need to gain a causal understanding of their behaviour and their circumstances (Coltheart 2017) and stories are what they use to make sense of themselves and the world around them (Bruner 2003). In *Interview*, the employer needs to justify her decision to the panel. In *Care Home*, the dementia patient needs to know where she is so that she can interact with the surrounding environment. In *Mask Mandate*, students need to explain why they are not following the rules. People's desired causal understanding often takes a narrative form and becomes an engaging story: we are all motivated to give our lives the shape of an easy-to-understand, ongoing narrative (Örulv and Hydén 2006), possibly with an upward trajectory (McAdams 2001) that indicates we are being true to ourselves and making progress.

Although stories often reflect people's biases and the limited information available to them, stories also enable people to connect with others by sharing meanings and themes from their lives (Stammers 2020) and to justify their behaviour to themselves and others, signalling that they are rational and trustworthy after all (Ganapini 2020).

What are stories evidence of?

As we saw, even the photo of a political candidate's childhood vacation in the mountains – when posted by his campaign team – and the photo of an artwork snapped during a museum visit – when the artwork depicts a civil-rights activist – can be *stories used as evidence*. The vacation shot is turned into evidence that the candidate is an authentic, well-rounded person who loves his homeland and is connected to his roots. The picture of the artwork becomes a reminder that the fight for equality needs to continue, inspired by the actions of the leaders and role models of the past.

There are many illustrations of the powerful role of stories as evidence in debates. The case of a teen in the United States who sought vaccination against his mother's will raised concerns about the reliance of anti-vaxxers on storytelling (Helmore 2019).

> Lindenberger said he had learned that his mother's unproven fears against vaccinations was received 'anecdotally', including a mistaken belief that the measles-mumps-rubella (MMR) vaccine raises the risk of autism in children, despite extensive scientific research that shows it does not.
> 'There's an important decision to be made between information provided,' Lindenberger said. 'Many people don't resonate well with data and numbers – they resonate better through stories.'
> We see that with the anti-vaccine community. A lot of the foundation they build with parents is on an anecdotal level, sharing stories and experiences. That speaks volumes to people because it reaffirms, especially for my mom, that her position is correct.

In the recent debate about the safety of COVID-19 vaccines, stories have been used to persuade participants of the debate of the risks of vaccination (by vaccine sceptics) but also of its benefits (by health authorities). Such stories, often including highly emotional content, seem to be taken in much higher regard and to have more significant effects on behaviour than other forms of evidence, such as diagrams comparing the risks of vaccination with the risks of catching COVID-19. This shows that stories are not just *a problem* for effective science communication – because they attract more attention than other data – but may also inspire a *solution*. By grabbing attention with a story, science communicators can encourage health-promoting behaviours.

It is now more readily recognised in the literature that stories often aim to persuade in the context of debating controversial issues and are generally successful at doing

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

so precisely because they do not have a traditional argumentative structure (Kubin et al 2021):

> The rich literature on political persuasion further highlights the ability for narratives to persuade – often because narratives typically present information 'peripherally,' minimizing the likelihood for counterarguments from 'central' processing.

For the authors, divulging personal experience of harm is the most effective way to persuade opponents to bridge disagreements and political divides, and gains storytellers respect to a greater extent than citing facts or articulating arguments.

In a recent discussion of governmental pro-vaccination campaigns at the time of COVID-19, stories are also considered more effective than other forms of information (Rogers 2021):

> Evidence suggests vaccine-hesitant groups are less likely to respond to factual information particularly from 'pro-vaccine' sources. But they may respond more to personal stories about the effects of the virus. In my area of research, we call these stories 'cultural health narratives'. Within the anti-vax movement, these narratives are often powerful stories of people negatively affected by vaccinations, or what they believe are vaccine-related side effects. These emotional accounts are very powerful because we're attracted to narratives and we live our lives through them.

Stories are also described as the way forward when attempting to persuade people to accept health interventions that are potentially risky (Rogers 2021):

> When we hear a story, we often lower our guard and tend to start responding emotionally to the characters. Parents, educators and religious leaders have long used this as a way of teaching. Governments could use storytelling to potentially improve COVID vaccination rates particularly among those who are unlikely to get the jab. Governments could add emotional health stories to their vaccination messages. These narratives could show the negative effects of the virus on people's lives, and/or they can be used to show the positive effects of vaccinations to help avoid disease.

In the climate change wars, the role of stories, when vividly presented by journalists, has been also acknowledged as a game changer. In a recent interview (Buonocore 2019), the author Dan Fagin says:

> [T]he journalists that are having the greatest impact on their work on climate change have chosen to tell true stories of people impacted, presenting them with empathy. Their work is building an understanding in the broader public that, whether you believe in climate change or not, the weather is changing, it is affecting people's lives, and we need to help these people. If journalists are able to work this way, they are more likely to make a difference and bring positive change. In my experience, the best climate writers are aware of this.
>
> We know that people do not retain dry information and data. It doesn't resonate with people emotionally. If we want people to remember something, and ultimately act on it, the content needs to come in the form of a story, in the form of narrative, with characters, drama and a connecting thread. Journalism needs all these things in order to be impactful and make a difference in people's lives.

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

If stories can both distract from and enhance science communication, what seems to be essential is to support the capacity for citizens to discriminate between 'good' and 'bad' stories, or stories that constitute good evidence from stories that do not. Although all stories are interesting and informative as 'windows into another world', only some stories are convincing evidence for a controversial position in a debate. The fact that stories are used in debates requires us to establish some criteria of evaluation for stories as evidence that can be easily applied.

That is too ambitious a goal for our present purposes, but here is a first step in what seems to us the right direction. In the context of a debate, we need to know what stories give us information about. In first-person accounts of people who describe what they are or what they did, the widespread phenomenon of confabulation suggests that the accounts may reflect how people see themselves, and how they want to be seen by others, and not how they actually are. In *Interview*, the employer wanted to present herself as a rational decision maker; in *Care Home*, the dementia patient wanted to project an image of independence and self-sufficiency; and in *Mask Mandate*, the students wanted to be seen as defenders of individual freedom against the overreach of powerful institutions.

More generally, first-person accounts of people who explain their behaviour are projections of what people see as the representative reasons behind their actions and choices, as opposed to an accurate causal account of the relevant factors leading to their behaviour. For instance, in *Interview*, the employer wants to present the choice of George as the outcome of a decision-making process based on the relevant evidence, as opposed to the outcome of a racist or sexist bias. In situations that are novel and puzzling, the story may be an attempt to provide a causal account of the events that is meaningful to the person given their previous experience or future goals. In *Care Home*, the dementia patient uses the general knowledge available to her to make sense of the features of her environment, rendered confusing by her memory impairment. And in *Mask Mandate*, it is the consumption of certain rhetoric on social media, a constant association between the refusal to comply with safety regulations and the love of freedom, that influences the students' understanding of their own choices.

The tendency to confabulate and the prevalence of the phenomenon do not render all stories 'bad' evidence by default. Rather, they show that we need to acknowledge the importance and value of people's sense of themselves as capable of intervening in their environment, as *agents*, in remembering, reporting, and constructing their own experiences. Pointing to the potential risks of confabulation does not mean that storytellers should be blamed for the fact that their stories are unreliable or ill-grounded. When people seek a causal understanding of their problems ('I cannot get an appointment with my doctor') or tragic events in their lives ('Baby Gromowski died'), they respond to significant setbacks and even tragedies by attempting to restore a sense of order and control that is central to their agency. They want the problem to be solved and the tragic event to never occur again. There is nothing blameworthy in seeking a causal explanation for those purposes – and the explanation is only as good as the information people can access and process. So, the difficulty in accessing healthcare is attributed to excessive immigration and the death of the baby is attributed to his recent immunisation. In this way, people place a short-term defeat in a bigger framework that enables them to cultivate some hope for the future: 'After Brexit, EU citizens won't be able to access healthcare services and my doctor will have more time for me'; 'Baby Gromowski is dead, but I will do all I can to prevent other babies from dying'.

Even when people confabulate, their stories can provide valuable information. Confabulations may not be the descriptions and explanations of the events supported by the best evidence, but they tell us something. By listening to people's stories, we learn about their values and appreciate what they care about and what motivates them.

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

This helps us see the world through their eyes, bringing to the fore issues that we might have dismissed otherwise, often from a position of ignorance or privilege. Just like the use of stories can be deployed in science communication or scaremongering – e.g., to educate people about vaccines or to persuade them that vaccines are unsafe – confabulatory explanations can have some use depending on how they are constructed, shared, and received. We can learn from stories involving confabulation. When used as an argument against vaccine safety, an account of how an immunisation in childhood resulted in a person's health problems may not teach us about the actual effects of vaccines or why pharmaceutical companies develop them. But it can teach us about that person's fears and values, and how they make sense of their own experiences. This can inform further interactions with the person who told the story and future attempts at communicating science to people who are likely to share the same fears and values.

How are debates taking place in a digital landscape affected by confabulation? As we saw, confabulation is partly driven by the motivation to order individual experiences into meaningful narratives and to build and reflect on identities that users share with others. Platforms on social media offer multiple ways to achieve those goals, providing a vast audience for people's stories – this is probably why people are so drawn to using those platforms. Because of the increased reach of digital storytelling, users consume many shared stories, stories that inspire them and resonate with them. What audiences may not realise is that some of those stories are used as evidence for controversial claims, and resources are needed to evaluate the causal connections that would make those stories good evidence for said claims. When such resources are unavailable, misinformation involving inaccurate causal explanations can be disseminated widely and quickly.

Conclusions

In personal stories, people make sense of significant events in their lives and project their image of themselves. Experiences are often remembered in such a way as to fit into the wider scheme of the person's goals and sense of purpose. These influences can become more powerful in people whose executive processes are compromised, because memory retrieval in clinical confabulation is more heavily driven by motivational factors and less effectively checked against other memories (Fotopoulou 2009, 278–280).

With personal stories, people do not merely share experiences but also argue for a particular, often controversial, viewpoint. Stories are informative whether or not they involve confabulation. Not only do they point to what the storyteller cares about, but they also tell us about what the storyteller values and what kind of person they aspire to be. But stories are not by themselves evidence for the viewpoints they are used to support. Brexit may not be the solution to the problems of national healthcare provision. Refusing immunisations may not prevent babies from dying. Independent evidence is needed for causal claims to be established.

In other words, stories as evidence for controversial claims need to be critically assessed, especially when they are shared on online platforms and have a greater reach and influence on public opinion. This might enable us to prevent misinformation and enhance the quality of debate among citizens. Acknowledging at the same time people's irresistible attraction to stories and their vulnerability to confabulation is instrumental to developing a satisfactory evaluative framework.

Acknowledgements. In the preparation of this article, the authors acknowledge the support and feedback of the Women in Philosophy group and the Belief and Delusion reading group at the University of Birmingham.

https://doi.org/10.1017/mem.2021.5 Published online by Cambridge University Press

Data availability statement. The current research article analyses and interprets data that have been previously published. References to the original articles have been provided throughout.

Funding. This work received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. L.B. and K.M.-H. declare none.

References

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 (2009) 'Faultless ignorance': Strengths and limitations of epistemic definitions of confabulation. *Consciousness & Cognition* **18**(4), 952–965.

Bortolotti L and Jefferson A (2019) The power of stories: Responsibility for the use of autobiographical stories in mental health debates. *Diametros* **60**, 18–33.

Bortolotti L and Sullivan-Bissett E (2018) Epistemic innocence of clinical memory distortions. *Mind & Language* **33**(3), 263–279.

Bruner J (2003) *Making Stories*. Cambridge, MA: Harvard University Press.

Buonocore M (2019) Storytelling is part of the solution to the climate dilemma—interview with Da Fagin. *Foresight*, 3 April. Available at https://www.climateforesight.eu/future-earth/storytelling-is-part-of-the-solution-to-the-climate-dilemma/ (accessed 1 July 2021).

Coltheart M (2017) Confabulation and conversation. *Cortex* **87**, 62–68.

Dayter D (2015) Small stories and extended narratives on Twitter. *Discourse, Context & Media* **10**, 19–26. doi:10.1016/j.dcm.2015.05.003.

Fotopoulou A (2009) Disentangling the motivational theories of confabulation. In Hirstein W (ed.), *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy*. New York: Oxford University Press, pp. 263–289.

Ganapini MB (2020) Confabulating reasons. *Topoi* **39**(1), 189–201.

Helmore E (2019) I defied mother to get vaccinated for safety of me and others, US teen says. *The Guardian*, 5 March. Available at https://www.theguardian.com/us-news/2019/mar/05/ethan-lindenberger-ohio-congress-vaccines (accessed 1 July 2021).

Hirstein W (2005) *Brain Fiction: Self-deception and the Riddle of Confabulation*. Cambridge: MIT Press.

Jenni K and Loewenstein G (1997) Explaining the identifiable victim effect. *Journal of Risk and Uncertainty* **14**(3), 235–257. doi:10.1023/A:1007740225484.

Kubin E, Puryear C, Schein C and Gray K (2021) Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences* **118**(6), e2008389118. doi:10.1073/pnas.2008389118.

Liebhart K and Bernhardt P (2017) Political storytelling on Instagram: Key aspects of Alexander Van der Bellen's successful 2016 presidential election campaign. *Media and Communication* **5**(4), 15–25. doi:10.17645/mac.v5i4.1062.

McAdams DP (2001) The psychology of life stories. *Review of General Psychology* **5**(2), 100–122.

O'Connor C and Weatherall JO (2019) *The Misinformation Age*. New Haven, CT: Yale University Press.

Örulv L and Hydén LC (2006) Confabulation: Sense-making, self-making and world-making in dementia. *Discourse Studies* **8**(5), 647–673.

Page R, Harper R and Frobenius M (2013) From small stories to networked narrative: The evolution of personal narratives in Facebook status updates. *Narrative Inquiry* **23**(1), 192–213.

Rogers M (2021) Why telling stories could be a more powerful way of convincing some people to take a COVID vaccine than just the facts. *The Conversation*, 15 February, 254–267. Available at https://theconversation.com/why-telling-stories-could-be-a-more-powerful-way-of-convincing-some-people-to-take-a-covid-vaccine-than-just-the-facts-155050 (accessed 1 July 2021).

Shelby A and Ernst K (2013) Story and science: How providers and parents can utilize storytelling to combat anti-vaccine misinformation. *Human Vaccines & Immunotherapeutics* **9**(8), 1795–1801. doi:10.4161/hv.24828.

Smith B (2021) Students protest state's school mask mandate. *CBSN Pittsburgh*, 7 September. https://pittsburgh.cbslocal.com/2021/09/07/greater-latrobe-high-school-face-mask-protest/ (accessed 1 July 2021).

Stammers S (2020) Confabulation, explanation, and the pursuit of resonant meaning. *Topoi* **39**(1), 177–187.

Stammers S and Bortolotti L (2020) When the personal becomes political: How do we fulfil our epistemic duties relative to the use of autobiographical stories in public debates?. In McCain K and Stapleford S (eds), *Epistemic Duties: New Arguments, New Angles*. New York: Routledge, pp. 254–267.

Sullivan-Bissett E (2015) Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition: An International Journal* **33**, 548–560. https://doi.org/10.1016/j.concog.2014.10.006

Villaespesa E and Wowkowych S (2020) Ephemeral storytelling with social media: Snapchat and Instagram stories at the Brooklyn Museum. *Social Media + Society* **6**(1), 1–13. doi:10.1177/2056305119898776.

Voytko L (2020) A timeline of Trump's misleading 9/11 claims. *Forbes*, 11 September. Available at https://www.forbes.com/sites/lisettevoytko/2020/09/11/a-timeline-of-trumps-misleading-911-claims/?sh=212d2d085836 (accessed 1 July 2021).

Kathleen Murphy-Hollies is a doctoral researcher at the University of Birmingham (UK), specialising in the area at the intersection of ethics and the philosophy of the cognitive sciences. Her doctoral project is about the impact of confabulation on how agents embody moral virtues.

Lisa Bortolotti is Professor of Philosophy at the University of Birmingham (UK) specialising in the philosophy of the cognitive sciences, philosophy of medicine, and issues in biomedical ethics. Her main interest is in the strengths and limitations of human agency and she has written extensively on delusion and confabulation.

Cite this article: Murphy-Hollies K, Bortolotti L (2022). Stories as evidence. *Memory, Mind & Media* **1**, e3, 1–12. https://doi.org/10.1017/mem.2021.5

Paper 3: Self-Regulation and Political Confabulation

Murphy-Hollies, K., 2022. Self-Regulation and Political Confabulation. *Royal Institute of Philosophy Supplements*, 92, pp.111-128, DOI: https://doi.org/10.1017/S1358246122000170.

Self-Regulation and Political Confabulation

KATHLEEN MURPHY-HOLLIES

Abstract
In this paper, I discuss the nature and consequences of confabulation about political opinions and behaviours. When people confabulate, they give reasons for their choices or behaviour which are ill-grounded and do not capture what really brought the behaviour about, but they do this with no intention to deceive and endorse their own accounts. I suggest that this can happen when people are asked why they voted a certain way, or support certain campaigns, and so on. Confabulating in these political contexts seems bad because we do not get a fully truthful account of why some political choice was made, and so the reasoning behind the choice is under-scrutinised. However, I argue that if people have a virtue of self-regulation, confabulation in political contexts can actually be part of the process of coming to better understand our political choices and embody more consistently the political values which we ascribe to.

When individuals make political decisions, such as voting for certain parties or supporting certain campaigns, they often give and receive explanations for why they are inclined to choose or behave in that way. Others who hear those explanations can find them implausible, unsatisfying and even frustrating. Imagine a woman who decides to vote for the UK to leave the European Union. When her peers ask her why she decided to vote in that way, she says that she doesn't like the power and influence from Europe in the UK and that it costs the UK too much money to be in the European Union. However, her peers are surprised to hear this because she did not seem to have paid any attention at all to these things before casting her vote. They suspect that other things, such as a sense of national pride, might have had a more powerful sway on her coming to vote for Brexit.

Explanations which are unsatisfying in this way may be frustrating because they are confabulations. When individuals confabulate, they don't track what has really brought about their decisions and actions, and they get something wrong about the world. Sometimes this means that harmful and irrational political decisions do not receive the scrutiny they deserve, from the individual giving the explanation or from others who receive it.

doi:10.1017/S1358246122000170

Individuals confabulate in many different contexts, whether it be consumer choice, moral convictions, or aesthetic preferences (Nisbett & Wilson, 1977; Hall et al, 2010; Haidt, 2001; Wilson et al, 1993). After elaborating on what confabulation is, in this essay I will explore confabulation in the context of politically charged decisions and behaviours, which have the potential to significantly affect the lives of many other people. Confabulations also often reflect self-concepts (i.e., traits and values which individuals ascribe to themselves) and identities (the type of person they see themselves overall as being).

It seems that confabulation is something we should try to avoid in political decision-making, but despite the problems it can cause, I will argue that confabulation can be an integral part of individuals coming to scrutinize their political actions further and make relevant changes to either their outlook or their political decision-making. Specifically, I suggest people will need the virtue of 'self-regulation' to do so. Exercising this virtue results in good management of the self, such that there is a good alignment between the descriptions someone gives of themselves and the reality of their behaviour. The virtue of self-regulation will consist in various attitudes such as being open-minded, curious, and receptive to the ideas of others about oneself.[1] With these attitudes in place, even confabulatory justifications of harmful behaviour can be conducive to the formation of behaviour which better aligns with values, in the long run. Both confabulation, recognised as a form of epistemic engagement and curiosity, and the virtue of self-regulation, as something which enables individuals to embody their professed values, can therefore be valuable.

I will focus on explanations which people gave for voting for significant political change, such as in voting for Brexit, or for Donald Trump, or in campaigning against lockdown restrictions during the Coronavirus pandemic.[2] I am interested in how confabulatory explanations made references to highly valued political ideals in order to justify

[1] The attitudes which I outline as contributing to the virtue of self-regulation draw on ideas by DeBruin and Strijbos (2019) about attitudes which make up what they term 'self-know-how'; a skillset which enables individuals to preserve first person authority of their self-ascriptions despite confabulation.

[2] Elsewhere, Lisa Bortolotti and I (2022) argue that justifications of certain behaviours during the global Covid-19 pandemic – particularly of behaviours which flouted the rules – were confabulatory and referred to distinctly political themes.

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

choices and behaviour. However, in these cases of confabulation, the explanations mask that the actions do not actually reflect and embody the values they refer to. This is not good for the agent, and if the value is a positive one, any harm incurred by not actually embodying the positive value can harm others and this harm is hidden. For example, in being mistaken about an action of theirs truly embodying compassion, that individual may inadvertently harm others.

1. What is confabulation?

Given how widely prevalent confabulation is, it is important to consider what influence it could be having in political contexts which are so often pressed with moral concern and expressions of one's identity. Confabulation has raised questions about the possibility and reliability of self-knowledge, and whether agents really are in any better a position to know their own mental states precisely because they are their own (Scaife, 2014). However, I think that the occurrence of confabulation highlights underlying features and motivations of how and why individuals give reasons for their behaviour, which are not necessarily bad. This will become clearer throughout my paper.

Confabulation was originally studied in the context of psychiatric disorders, where it is often seen as a component of a number of conditions such as Alzheimer's, Anosognosia, or Korsakoff's syndrome (Hirstein, 2005). However, I am focusing on less severe instances of confabulation which occur in everyday situations with healthy individuals. When people are asked why they have some belief or why they have behaved a certain way, they may confabulate if they do not actually have an accurate answer (Sullivan-Bissett, 2015, p. 552). This is the crux and mystery of confabulation; instead of their attention being drawn to their lack of explanation, individuals produce an explanation which they do sincerely believe. However, these explanations are 'ill-grounded' (Bortolotti, 2018, p. 237). This means that their answers are not based on the relevant evidence and fail to capture what actually brought that behaviour about. Their answers often include false statements about the world.

Crucially, confabulators do not realise that their confabulations are ill-grounded in this way, and that they can't be giving an accurate account of why they came to behave in the way that they did. There's no intention to deceive, as they fully endorse the account they put forward, thinking that it is the truth. This is partly because they do not have access to, or don't know of, more accurate

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

explanations of their behaviour (Bortolotti, 2020, p. 15). This might be because some cognitive processes operate at the sub-conscious level and then individuals cannot know through introspection about the source of, for example, their intuitions or gut feelings, or of subtle external influences on their thinking and behaviour. This is the kind of 'gap' which can be filled with confabulation.

For example, in one study, participants read a vignette about a brother and sister who decide to sleep together, and are then asked to give reasons why the act was morally right or wrong (Haidt, 2001). Participants often gave reasons that cited factors which the vignette made clear did not apply; the siblings used contraception and their relationship was not affected, and so concerns about pregnancy and damage to the relationship were not applicable and could not have driven the moral judgement. When participants are reminded that these concerns do not apply, they continue to believe that the act was wrong nevertheless and are uncomfortably 'dumbfounded' as to why. Haidt suggests that these reasons were produced post-hoc, after a strong initial gut feeling of moral disgust. This disgust is what actually drove their decision, rather than the moral reasoning offered (ibid, p. 815).

Why do people do this? There are various possible motivations. Individuals do not want to be without an answer, unable to explain themselves and embarrassingly dumbfounded (Sullivan-Bissett, 2015, p. 552). But more generally, individuals want to provide an explanation which, although it may not be strictly accurate, meets other needs for them. They are motivated by the need to have a causal understanding of themselves and the circumstances they are in (Coltheart, 2017). They want to signal to other people that they are rational, competent, and trustworthy, (Ganapini, 2020) and so they are motivated to provide explanations which present themselves positively and protect positive self-concepts (Sullivan-Bissett, 2015, p. 552). Other, perhaps more accurate explanations such as 'I just felt like it' or 'I did it at random' put them at risk of looking foolish or unkind. So, individuals want to provide explanations which paint good pictures of themselves.

Confabulations can serve as explanations and justifications, and they often take a narrative form. These things have powerful benefits for people; weaving experiences and behaviour into a sense-making story or narrative helps people understand them and give them meaning (Örulv and Hydén, 2006). And then, these narratives facilitate social communication and the sharing of themes and values which are important to them (Stammers, 2018). Individuals might be particularly keen to fulfil these needs when it comes to moral

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

issues, as is seen in Haidt's study. Individuals feel strongly about these topics and want other people to know that they are good, trustworthy, and share the same values.

Overall, I suggest that we can characterise something as confabulation if the explanation has the following features:

(1) It is produced post-hoc and the reasons given are not the ones which truly led the individual to act in that way.

(2) It is a false or ill-grounded explanation, in that it says something false about the outside world or, in some circumstances, is technically correct but only by chance.

(3) It is motivated in the ways discussed above; to paint a certain picture of the agent, and/or to create and communicate what is meaningful to them.

2. Politically charged confabulation.

I suggest that individuals will have a tendency to confabulate in explaining and justifying politically charged decisions and behaviours. Our political views can be close to our hearts, forming significant parts of our identities and relating closely to our moral views. Political choices affect not only ourselves and the people around us, but how we are represented on a global stage and work alongside other global bodies and powers. This motivates giving explanations, particularly ones that make us look rational and good, even if they do not accurately reflect the facts of the matter.

A number of tumultuous events recently have brought this out. Britain unexpectedly voted to leave the European Union in 2016, Donald Trump was controversially elected president of the USA in the same year, and in 2020 world governments had to navigate a global pandemic owing to the Covid-19 virus. Each of these events brought our moral convictions to the surface, as the consequences were highly significant for huge numbers of people. Individuals were keen to be able to justify their own decisions and to demand explanations from others for their choices. All these events were also plagued with issues of 'fake news' and 'alternative facts' being circulated, causing confusion, and making it difficult to know the facts of the matter. Sometimes this could provide more convenient versions of the 'truth'.

Return to the example of the woman explaining her decision to vote for Brexit, whose statements are found unconvincing. Perhaps the woman in question is generally quite patriotic, feeling particularly proud and protective of her country and its values. She is therefore

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

drawn to the idea of Brexit through gut feeling, but she is not fully aware of the source of this inclination. After she votes for Brexit, she is asked by her friends why she did so. She doesn't want to seem politically ignorant or to have made such an important political decision flippantly, so she thinks about all the campaigning she saw for Vote Leave even though she didn't actually pay much attention to it before now. She recites that voting leave secures freedom from European lawmakers, means that the UK will keep more money, and will curb rates of immigration. She is confident that these are the reasons why she voted for Brexit.

I suggest that this kind of response is confabulatory; the reasons she gives for voting for Brexit never particularly crossed her mind in the run up to voting to leave, so the explanation is post-hoc. Her answer doesn't capture that her patriotic gut feelings and negative intuitions about European influence were factors in her coming to vote for Brexit, and so has feature (1) from the previous section.

We can see this kind of mechanism at play in considering the controversial case of extensive targeted advertising being used to show pro-Brexit advertisements to millions of people on facebook (BBC 2018). This could have had a significant impact on people's choices and the outcome of the referendum, yet this influence likely stemmed from merely an increased sense of familiarity with basic pro-Brexit sentiments and values. This could have contributed to, for example, our imagined voter's unexamined nationalistic gut feelings. But importantly, this is not a carefully considered change in decision-making. It is unlikely that the individuals would have endorsed this intervention on their decision-making, and agents are more likely to produce confabulatory explanations such as the one our voter gives, than acknowledge the role of persistent advertising on their decision. In overlooking the influence of these more basic feelings of familiarity and intuition in explanations, and instead citing more complex political and economic reasons for their decisions, confabulators are wrong about the psychological processes which led to their decisions.

The explanation also has feature (2); the explanation makes demonstrably false statements about the world. Specifically, her answer draws on false statements about the reality and scale of immigration. Rates of immigration are widely misconceived in the UK, with people guessing that the proportion of immigrants in the population is twice as high as it actually is (Ipsos Mori, 2014). Her answer also draws on false statements about how much money the UK sends to the EU, how it is spent, and how it may in fact be received back, facts that were difficult to ascertain even for experts. Yet, our

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

Brexit voter doesn't convey this in her simple assertion that the UK 'pays too much money' to the EU. It is also false to say that she paid considerable attention to these political and economic factors, deliberating about them *before* coming to make her choice. It is important to note that these are not incorrect beliefs which she acts on; she acts on her nationalistic intuition and draws on these mistaken beliefs in her confabulation because they make her decision seem more rational and acceptable.

I have used an example of voting for Brexit because it is an example of a politically charged explanation which is likely to have particularly high pressures to preserve positive self-concepts, as described in feature (3). This is because, akin to going against government-mandated lockdown regulations and voting for Trump, these were political behaviours which aroused particularly strong reactions and were votes for substantial change rather than for keeping more established political arrangements. Therefore, our voter is subject to particularly stark pressures to provide an explanation which justifies (in her eyes) her actions, meeting feature (3). She produced this explanation for her choice because she wanted to be able to share and communicate her values relating to Britain and Brexit with her friends when they asked, and not to seem uneducated on the issue or that she didn't actually have a particularly firm reason for voting as she did. One way that she makes her decision particularly acceptable and appealing to others is by drawing on a valued cultural ideal. Specifically, the value of freedom, particularly from European influence and law-making, was one such valued cultural ideal which she drew on and ascribed to herself. Anti-lockdown movements and marking the end of lockdown as 'freedom day' similarly drew on the same value, and so confabulatory explanations for rejecting measures to contain Covid-19 could work in the same way, emphasising a cultural and personal value for the individual at the expense of acknowledging scientific data and expert opinion on the spread of coronavirus.

This is not to say that explanations for voting in other directions (for Remain, for the Democrats etc.) could not also draw on rousing personal and political values which might not be strictly relevant, as nearly all political decision-making will bring pressures to confabulate. But the cases of Brexit and rejecting lockdown also provide a clear example of drawing on similar political values. For instance, the day the UK left the European Union (31 January 2020) and the final day of UK lockdown (19 July 2021) were both heralded as 'Freedom Day' (Duffield, 2021; Honeycombe-Foster, 2019). The political ideal of freedom is highly valued in the UK,

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

and so it is powerful in glossing these political choices with some good-looking purpose.

3. Harms and costs of confabulation.

We are now in the position where an imperfect, confabulatory explanation has been offered, and the individual seems worse off for it. She is unaware of not having had a particularly well-formed reason for her decision and does not seem to be prompted to check whether her conceptions about, for example, immigration, are accurate.

The first type of concern relates to feature (1) of confabulations, when agents are wrong about the reasons they acted in some way and sometimes falsely posit certain psychological processes as being behind a decision. One reason we might care about agents 'filling in this gap' with an inaccurate explanation is that highly valued ideals get mis-applied, with consequences for others. For example, the decision to call the final day of the UK lockdown 'freedom day' was criticised for being directly at odds with the reality of the situation for many. For example, vulnerable individuals and young people who were not able to be vaccinated had to give up many freedoms in order to shield, and precipitating a fourth lockdown would reduce freedoms, and finally overloading the NHS to the point of barely functioning reduces freedoms (Ahuja, 2021).

It can sometimes seem that an individual's political choices bring about states of affairs which do not necessarily reflect that value very much. But, in their confabulations individuals draw on these highly valued political ideals and nevertheless ascribe them to themselves because it justifies the behaviour for them and satisfies other motivations in play described in feature (3); to have an explanation and understanding of one's decisions. We would want requests to explain and justify these decisions to be met with earnest thought and scrutiny, but this seems to be lost when a confabulation is offered instead. This fits the behaviour into a story that justifies it, protecting the reputation and self-concepts of the individual, and even imbuing it with a noble personal and political value.

A second type of concern arising from confabulation relates to feature (2) of confabulations; making ill-grounded claims. In drawing on false statements about the world, by fitting these false statements into narratives which make the decision more acceptable and rational, confabulation discourages the agent from thinking further about the possible falsity of those claims. Inaccurate claims about the nature of coronavirus put other people's health and lives

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

at risk, as well as the wider project of containing the virus. Yet these false claims will be reinforced for a person if they are propping up a confabulatory explanation which holds meaning for them.

4. The role of confabulation in self-regulation.

So far, I have described what confabulation in political contexts looks like and why we might not welcome it. In the rest of this essay I will argue that political confabulation can be beneficial alongside a virtue of self-regulation.

I have already touched upon some of the motivations which underlie confabulation; people want to have an understanding, to display certain concepts of themselves, to construct meaningful narratives of their experiences, and to communicate them with others. These are very worthwhile endeavours which bring great benefits to people. Not only is it psychologically valuable to be able to organise our experiences into a coherent story which makes sense, but individuals can use these crafted personal identities and senses of themselves to actually guide their future behaviours, making them a reality (Velleman, 2006). Bortolotti suggests that confabulation can even mean that in the long-run, agents end up with *more* accurate beliefs about themselves and the world. This is because in giving reasons, even though they may not be accurate, their sense of agency is emboldened, their self-concepts are enhanced, and they engage with peers (2018, pp. 239–40). This 'active engagement with the world is also an epistemic goal' (2018, p. 241). Agents can receive feedback from others on their explanations and begin to think more explicitly about their behaviour and what is driving it. So, there is not anything distinctly bad about the motivations which individuals have for confabulating. They are primarily for understanding, for meaningfulness, and for connectedness with others.

However, there seem to be frustrating cases of confabulation in which someone re-casts themselves and the situation again and again in a way which preserves their good self-image, even when it's not deserved. In these kinds of cases, a good picture of the person continues, whilst the harmful and misguided reality of their behaviour continues too. There's an uncomfortable gap between what they *do* and what they *say*. Ideally, we want people to be able to bridge this gap; to give more accurate explanations of what is driving their behaviour, or for their behaviour to more reliably and consistently reflect an appreciation and adherence to these values. If every time someone is confronted for not wearing a mask they

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

passionately espouse the value of personal freedom, they might only become more entrenched into that narrative despite the fact that they actually just couldn't be bothered to wear a mask and are indirectly curbing the personal freedoms of others.

So, while this is far from guaranteed, confabulation has the potential to impress valued self-concepts more explicitly into one's own mind and guide future behaviour to reflect those concepts and values.

5. The virtue of self-regulation.

The virtue of self-regulation will function like Cathy Mason's virtue of hope as described in this collection; as a structural virtue (2022). That is, rather than being directly related to a motivation for perceiving and responding correctly to certain values in the world, self-regulation is a valuable form of self-government which is called for in certain kinds of situations.

The situations which will call for the virtue of self-regulation are ones in which someone faces a possible inconsistency between their self-ascriptions and their behaviour, which is likely to happen when someone confabulates. Self-regulation is the matter of being able to effectively re-align these inconsistencies.[3] I suggest that someone with the virtue of self-regulation will be: open-minded about what may in fact be influencing their behaviour which is not being captured in confabulatory justifications; will be curious and attentive towards their own feelings, motivations and what may be causing them; will be receptive to the feedback received from others on the accuracy of their self-ascriptions, and will have the right amount of confidence in self-ascriptions such that they are not defended even in the face of compelling evidence to the contrary but neither do they crumble under the slightest pressure.

These attitudes are complex in that they equip the individual with the disposition to behave in the relevant ways when required (for

[3] Leon DeBruin and Derek Strijbos (2019) suggest that in order to close this gap, it helps if people have certain self-directed attitudes that make up a sort of skilful 'know-how' (as opposed to propositional knowledge of self-related facts). I also draw on their idea that how one goes about closing the gap is more important (and what we value more in others) than the gap itself, whether this be a large gap, as in cases of people with poor self-regulation, or a very small gap. They even describe *perfect* alignment as possibly indicating psychological, neurocognitive or personality vulnerabilities and an obsession with rigidity (p. 159).

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

example, actually changing one's mind when one encounters new evidence) and could be described as virtuous given that they reflect the virtue of self-regulation. They give the agent a mastery in managing their own behaviour given that it is in fact *their own* behaviour. Individuals do not simply 'read off' what their traits are from observing their own behaviour. Instead, in this process, they play an active role in shaping and negotiating their own behaviour and self-ascriptions.[4] Ideally, we want to be able to describe ourselves accurately but also live up to our own descriptions. This will involve dealing well with discrepancies between the two when they arise by adjusting either the ascriptions or the behaviour.

As with many virtues, the virtue of self-regulation will be subject to ideally operating within a 'golden mean'. Someone with a deficiency in courage might avoid dangers, or minimize and misrepresent dangers, and an excess of courage might lead someone to be reckless or to seek out confrontations. Similarly, a number of different attitudes and behaviours may come out of an excess or deficiency of self-regulation. In excess, individuals will be too preoccupied with the alignment of their self-ascriptions and behaviour to ever try anything slightly new, spontaneous, or 'out of character' for them. They would stringently avoid any self-ascriptions which could possibly be a little wishful or incorporate aspects of their idealised selves, which means that they lose the opportunity to regulate themselves 'up' to this desired self-image (as is suggested and described in Jefferson 2020). In deficiency, individuals live in a state of fantasy that they are already whatever person they perceive themselves to be, just in virtue of perceiving or desiring that they are that way. This could be completely at odds with reality. They will have no open-mindedness, no curiosity, and be utterly dismissive of the views and ideas of others with regards to their self-ascriptions and behaviour. Moreover, even in the right quantity, self-regulation could allow its possessor to be, for example, consistently unkind, impatient, or, in the political realm, fascistic. So, the virtue of self-regulation will also be subject to proper exercise in accordance with a more general ethical wisdom, or phronesis.

Finally, self-regulation also works as a corrective in Philippa Foot's sense (2002). That is, it prevents individuals from falling into the temptation of simply believing that they in fact are the type of person they admire and wish to be, simply because of their appreciation for that type of person. Generally, the appeal of being a good,

[4] For more on this 'regulative' dimension of self-knowledge, see McGreer (2015).

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

competent person is not difficult to find in people. The harder work of ensuring that they live up to such an image, avoiding the lures of self-deception and looking the other way when they do not live up to their ideals, is what takes some correcting.

In instances of political confabulation, individuals espouse certain political values and ascribe them to themselves. Then, if someone has the virtue of self-regulation, they are able to make the most of those self-ascriptions by more effectively and consistently embodying those values in their future decision-making and behaviour.

So, although we may have individual instances of confabulation, looking at a longer time-span, these inaccurate explanations can still have a role to play in bringing individuals to live up to the ascribed self-ascriptions. Without self-regulation and the attitudes described therein, individuals just continue to confabulate in all instances in which their behaviour does not reflect what they want it to. This is because they are unable to manage the misalignment; they can't consider other possible explanations for their behaviour, or integrate feedback from others about the plausibility of their offered explanations, or pay more attention to their thoughts and feelings. In the next section, I describe in more detail how this virtue and its associated attitudes could bring our Brexit voter to better embody the values she is espousing in her confabulatory explanation of why she voted for Brexit.

6. Confabulation and self-regulation at work.

Returning to our imagined Brexit voter, recall that she has provided a confabulatory explanation of her behaviour and the concern is that this has stopped her considering more earnestly what her reasons for voting for Brexit really were, and that sometimes harmful behaviour ends up being justified in this way.

I argue that this non-ideal, confabulatory explanation is better than being dumbfounded in that it affords the individual certain opportunities. Importantly, being dumbfounded is the only other option available to the would-be confabulator, because agents do not have access to more accurate explanations (Sullivan-Bissett, 2015, p. 552). Our Brexit voter does not have introspective awareness of the influence of her patriotic gut feelings on her decision to vote in that way. But her confabulation does make more explicit to herself and others the factors – and their associated values – which she wants and expects to be the causes of her behaviour. She signals that she values British independence and self-governance. Once she

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

expresses these self-concepts of being protective of Britain, she can then be more mindful going forward of whether her decisions reflect those values. Thus her self-ascriptions of value and her behaviour can become better aligned.

How well individuals do this will depend on whether they have the virtue of self-regulation. It is hard to imagine someone without any of the attitudes involved in this virtue (curiosity, openness to others' interpretations, attentiveness to feelings) not confabulating (and instead being dumbfounded) because they both reflect an epistemic curiosity to understand oneself and the circumstances. One of the first things an individual with a curiosity and open-mindedness for explanations of themselves is going to do, is to come up with a possible – if unideal – explanation. They are likely only to be more comfortable with being dumbfounded when it comes to very neutral topics, which they don't regard as being particularly reflective of their values and personalities more widely. I think this is unlikely to be the case with political decisions and actions.

Without the virtue of self-regulation, confabulation carries important costs. For example, in these cases, it could drive entrenchment of the attitudes that actually motivated the action and prevent further reflection. At worst, confabulation is not only a missed opportunity for individuals to present themselves with self-ascriptions which they can work with over time, but could lead to them being perceived by others as unreliable describers of themselves. DeBruin and Strijbos describe this as the most significant possible cost for agents with regards to confabulation (2019, p. 154). If this happens, agents are not considered reliable, trustworthy and intelligible members of the community. However, the person who is serially dumbfounded has no epistemic curiosity with regards to their behaviour at all, and so is also unlikely to reflect on their explanations and to experience some lack of intelligibility and isolation from others. (Imagine our Brexit voter looking bored and saying that she has 'no idea' why she voted for Brexit, when she is asked.)

In summary, we should not immediately judge an occurrence of confabulation as a wholly negative thing. It stems from good and epistemically valuable motivations. By this I mean that these motivations can improve the state of an individual's knowledge about themselves and the world. What is even better is if, further than this, agents have the virtue of self-regulation and can navigate the resulting feedback well and negotiate competently between their ascriptions and their behaviour. In the next section, I illustrate this point by showing how this could happen in the context of navigating 'undermining propaganda'.

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

7. Navigating Undermining Propaganda.

There are some political phenomena in which there are particularly stark 'gaps' or 'mismatches' between the values being espoused and the reality of that political goal being realised. This is exactly what we see in cases of 'undermining propaganda', so here I consider how the virtue of self-regulation and its associated attitudes can help individuals navigate such propaganda.

Jason Stanley's (2015) influential account of propaganda includes the following sub-type of propaganda:

> *Undermining Propaganda*: A contribution to public discourse that is presented as an embodiment of certain ideals, yet is of a kind that tends to erode those very ideals. Undermining propaganda involves a kind of contradiction between ideal and goal. It's an argument that appeals to an ideal to draw support, in the service of a goal that tends to erode the realization of that ideal. (2015, p. 53).

The ideals in questions can be any sort of ideal, but Stanley's focus is, usefully, on political ideals. He cites a number of examples of undermining propaganda, one of which being the 'war on drugs' in America in the 1980s-90s. The 'war on drugs' invoked the ideals of justice, law and order, and fair sentencing in a campaign which in reality exacerbated sentencing disparities for white and black people using cocaine and crack cocaine (ibid, pp. 59-60).

When we consider for ourselves which political messages we want to adhere to and which values we wish to embody, the attitudes involved in self-regulation can be beneficial in helping combat the effects of undermining propaganda by improving alignment between our espoused political values and the realities of our behaviour. We may be better able to *spot* undermining propaganda, and not personally participate in furthering its goals by following its message and thereby eroding the value we initially espoused. The contradiction between the ideal expressed in a political message and the realisation of that message is masked by individuals having flawed ideological beliefs (ibid, p. 57). If these ideologies are accepted because there is something personally appealing about them, or in other words because the individual feels that accepting them exemplifies some trait that they (gladly) perceive themselves as having, then having the virtue of self-regulation would put pressure on the link between one's acceptance of this ideology and the reality of their behaviour responding to the propaganda.

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

Here is an example. Imagine that Jack and Jill are both fed up with the state of American politics. They believe that politicians are out of touch and haven't made any changes which actually improve their lives in any way, because politicians are too caught up in various scandals and alliance-forging which make up political life. So they both have a general, unexamined sense of disenfranchisement, and this is what brings them to decide to vote for Donald Trump. Then, they are both drawn to Trump's presidential campaign messaging which invokes the values of anti-corruption ('drain the swamp'), law and order, and politics re-centering around the average American. In confabulatory explanations of why they voted for Trump, they cite an adherence to and having been moved by these values. However, this isn't a completely accurate explanation because adherence to these values isn't what drove their sense of disenfranchisement; in fact it is the other way around. Their basic sense of disenfranchisement drove their decision and then these values of law and order are drawn on post hoc in order to justify this political choice in a way which is more satisfactory to them. It has meaning for them and allows them to communicate with others that they appreciate these political values.

Over time it becomes clear that Trump appoints his friends and family members to key positions in the White House, engages in unlawful tax-dodging and lives a life very unlike the 'average American'. This points to Trump's election campaign being an instance of undermining propaganda because electing him did not further those values. It also raises issues about the possible falsity of claims about Trump's actions and practices, such as whether he engages in tax-dodging. Yet, these claims are unexamined when they become a part of Jack and Jill's confabulations about why they voted for Trump.

Jack not only subscribes to these political values, but in his eyes they support his conception of himself as being independent, self-sufficient, and able to take care of himself. Also, that he is savvy to the constant deceit and deal-making amongst politicians, and this explains for him why his general situation and quality of life never improves. He is too confident in his own appraisal of what is wrong with politics and receives psychological comfort from it because it satisfies his need for an understanding of the world. He has too little interest and respect for the opposing thoughts and experiences of those around him. He isn't curious enough and is too inflexible in his thinking to consider whether Trump being president really furthers the de-corruption of politics, and whether his illegal tax-dodging is actually an endorsement of breaking the law when it benefits you. Jack does not have the virtue of self-regulation and

https://doi.org/10.1017/S1358246122000170 Published online by Cambridge University Press

his confabulations about voting for Trump will include ill-grounded statements which prop up a narrative of having made a political choice which embodies his values, and these statements are unlikely to be scrutinised further.

Jill, on the other hand, despite self-ascribing the same values of being against political corruption and wanting average Americans to be more central to political decisions, is more open to the perspectives of others in considering what a Trump presidency would look like and mean for them. In talking through with peers her explanation as to why she voted for Trump, she comes to see how angry and disenfranchised she feels about American politics and she has an interest and inquisitiveness as to how and where that anger is best directed. Her self-confidence is not so low that she quickly assumes she is wrong about something if she hears an opposing view, but when some friends ask her whether avoiding paying your taxes really reflects her value of being law-abiding and understanding the lives of 'average Americans', she reflects on this. There is a possibility that she comes to stop seeing Trump and his presidential campaign as something that will truly further the values of anti-corruption, law and order, and a focus on the 'average American'.

Jill therefore demonstrates how having the virtue of self-regulation can mean that her confabulating actually played an integral part in her coming to consider whether a political action of hers – voting for Trump – actually aligned with her political values. In this turbulent political world, the drawing together of one's political outlook and political decision-making, unhampered by misleading political propaganda, is particularly valuable.[5]

University of Birmingham
klm276@student.bham.ac.uk

[5] I would like to thank Lisa Bortolotti, Iain Law and Quassim Cassam for extensive and thoughtful comments on earlier drafts of this paper. Also, to the Women in Philosophy group at the University of Birmingham for listening to these ideas and providing invaluable feedback. Finally, to the editors Anneli Jefferson and Jonathan Webber, whose responses and suggestions I am incredibly grateful for and made this paper substantially better. Many of the other contributors to this volume also helped me better formulate these ideas.

References

A. Ahuja, 'Monday is surrender day, not freedom day', Financial Times, https://www.ft.com/content/c9a6c0f0-985c-4563-91bb-aee51f0ab926. (2021) Accessed July 2021.

BBC News, 'Vote Leave's targeted Facebook ads released by Facebook', BBC News, https://www.bbc.co.uk/news/uk-politics-44966969. (2018) Accessed September 2021.

L. Bortolotti, 'Stranger than fiction: costs and benefits of everyday confabulation', in *Review of philosophy and psychology*, 9 (2018) 227-249.

L. Bortolotti, *The epistemic innocence of irrational beliefs*, (Oxford University Press, 2020).

L. Bortolotti and K. Murphy-Hollies, 'Exceptionalism at the time of COVID-19: where nationalism meets irrationality' in *Danish Yearbook of Philosophy* (2022), 1–22.

M. Coltheart, 'Confabulation and conversation', *Cortex* 87 (2017) 62–68.

L. De Bruin and D. Strijbos, 'Does Confabulation Pose a Threat to First-Person Authority? Mindshaping, Self-Regulation and the Importance of Self-Know-How', *Topoi*, 39 (2019) 151–161.

C. Duffield, ''Freedom Day' in pictures: England marks easing of lockdown restrictions on amid rising covid cases', i news, https://inews.co.uk/news/uk/freedom-day-pictures-england-lockdown-easing-19-july-rising-covid-cases-1110556. (2021) Accessed July 2021.

P. Foot, *Virtues and Vices,* (Oxford: Oxford University Press, 2002).

M. B. Ganapini, 'Confabulating reasons', *Topoi*, 39 (2020) 189-201.

J. Haidt, 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement', *Psychological Review*, 108 (2001) 814–834.

L. Hall, P. Johansson, B. Tärning, S. Sikström, and T. Deutgen, 'Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea', in *Cognition*, 117 (2010) 54–61.

W. Hirstein, *Brain fiction: Self-deception and the riddle of confabulation,* (MIT Press: USA, 2005).

M. Honeycombe-Foster, 'MPs in fresh demand for Big Ben chime to mark Brexit 'Freedom' day', Politics Home, https://www.politicshome.com/news/article/mps-in-fresh-demand-for-big-ben-chime-to-mark-brexit-freedom-day. (2019) Accessed July 2021.

Ipsos Mori, 'Perceptions are not reality: Things the world gets wrong', https://www.ipsos.com/ipsos-mori/en-uk/perceptions-

are-not-reality-things-world-gets-wrong. (2014) Accessed July 2021.
A. Jefferson, 'Confabulation, rationalisation and morality', *Topoi*, 39 (2020) 219–227.
B. Johnson, PM speech in Greenwich: 3 February 2020, https://www.gov.uk/government/speeches/pm-speech-in-greenwich-3-february-2020. (2020) Accessed July 2021.
McGeer, 'Mind-making practices: the social infrastructure of self-knowing agency and responsibility', *Philosophical Explorations*, 18 (2015) 259–281.
R. E. Nisbett and T. D. Wilson, 'Telling More Than We Can Know: Verbal Reports on Mental Processes', in *Psychological Review*, 84 (1977) 231–259
L. Örulv, and L. C. Hydén, 'Confabulation: Sense-making, self-making and world-making in dementia', in *Discourse Studies*, 8 (2006) 647–673.
Rev Transcript, 'Donald Trump Tours the Ford Plant Without a Mask, Explains Why', Rev, https://www.rev.com/blog/transcripts/donald-trump-tours-the-ford-plant-without-a-mask-explains-why. (2020) Accessed July 2021.
R. Scaife, 'A problem for self-knowledge: the implications of taking confabulation seriously', in *Acta Analytica*, 29 (2014) 469–485
S. Stammers, 'Confabulation, explanation, and the pursuit of resonant meaning', in *Topoi*, 39 (2020) 177–187
J. Stanley, *How propaganda works*, (Princeton University Press, 2015).
D. Strijbos and L. de Bruin, 'Self-interpretation as first-person mindshaping: Implications for confabulation research', *Ethical Theory and Moral Practice*, 18 (2015) 297–307.
E. Sullivan-Bissett, 'Implicit bias, confabulation, and epistemic innocence', in *Consciousness and Cognition*, 33 (2015) 548–560.
J. D. Velleman, 'The Self as Narrator', in *Self to Self: Selected essays*, (USA, Cambridge University Press, 2006).
Yahoo News, 'Florida's Anti-Maskers Are Taking A Stand', *Yahoo News*, https://news.yahoo.com/floridas-anti-maskers-taking-stand-231500975.html. (2020) Accessed July 2021.
T. D Wilson, D. J. Lisle, J.W. Schooler, S.D. Hodges, K.J. Klaaren and S.J. LaFleur, 'Introspecting about reasons can reduce post-choice satisfaction', in *Personality and Social Psychology Bulletin*, 19 (1993) 331–39.

Paper 4: Exceptionalism at the time of COVID-19: where nationalism meets irrationality

Bortolotti, L. and Murphy-Hollies, K. (2022), 'Exceptionalism at the time of COVID-19: where nationalism meets irrationality', in the *Danish Yearbook of Philosophy*, 55(2), pp.90-111. DOI: https://doi.org/10.1163/24689300-bja10025

BRILL

DANISH YEARBOOK OF PHILOSOPHY 55 (2022) 90–111

DANISH YEARBOOK OF PHILOSOPHY

brill.com/dyp

Exceptionalism at the Time of COVID-19: Where Nationalism Meets Irrationality

Lisa Bortolotti | ORCID: 0000-0003-0507-4650
Philosophy Department, University of Birmingham, Edgbaston, UK
l.bortolotti@bham.ac.uk

Kathleen Murphy-Hollies | ORCID: 0000-0001-5061-4674
Philosophy Department, University of Birmingham, Edgbaston, UK
KLM276@student.bham.ac.uk

Abstract

Exceptionalism is the view that one group is better than other groups and, by virtue of its alleged superiority, is not subject to the same constraints. Here we identify national exceptionalism in the responses made by political leaders in the United States and the United Kingdom to the COVID-19 pandemic in early 2020. First, we observe that responses appealed to national values and national character and were marked by a denial of the severity of the situation. Second, we suggest an analogy between national exceptionalism and unrealistic optimism, i.e., people's tendency to make rosier predictions about their future than is warranted by the evidence due to illusions of superiority and control. Finally, we argue that, at the national level, exceptionalism gave rise to an assumption of invulnerability that made for slow responses to the pandemic, and at the individual level, it served as a justification of people's failures to adopt safety behaviors.

Keywords

COVID-19 – unrealistic optimism – national values – exceptionalism – national character – freedom – positive illusions – confabulation

1 Exceptionalism and the Pandemic

Broadly, exceptionalism is the idea that one group is superior to others. More specifically, it is the view that, due to its alleged superiority, a certain group is not subject to the same constraints as other groups and deserves special treatment.[1] Exceptionalism has been advocated in various contexts, such as the relationship between human and nonhuman animals, where the human species is the one regarded as superior. Here we focus on *national exceptionalism* and argue that there are some interesting connections between this phenomenon and the responses of some political leaders to the COVID-19 pandemic in early 2020.

In particular, we argue that the combination of exceptionalism and nationalism can give rise to an assumption of invulnerability that makes for slow and ineffective responses to threats and justifies failures to comply with safety behavior at the individual level. There is no better example of the role of nationalist rhetoric in political decision-making than the initial responses of countries such as the United Kingdom and the United States to the COVID-19 global pandemic.[2] In this paper, we focus on these two examples because the prominence of those nations within the international community means that the behavior of their political leaders was more influential and more closely scrutinized, even though other countries, such as Brazil, Sweden, and Denmark, were charged with exceptionalism at the time. In the case of the UK and the US, more evidence is available of the explicit justifications provided for the policies adopted by their leaders and of the reactions to such justifications, nationally and internationally.

In sections 2 and 3, we review how political leaders in the UK and the US openly advocated the superiority of the British and American national character and appealed to national values in their initial responses to COVID-19 in early 2020, based on transcripts of their speeches, newspaper articles, and

1 On exceptionalism at the national level, see John A. Agnew, "An Excess of 'National Exceptionalism': Towards a New Political Geography of American Foreign Policy," *Political Geography Quarterly* 2, no. 2 (1983): 151–166.

2 Other countries have been described as exhibiting exceptionalism—for instance Sweden, Denmark, and Brazil—but we will not consider them here. For more information about claims of exceptionalism as they apply to Sweden, see Staffan Andersson and Nicholas Aylott, "Sweden and Coronavirus: Unexceptional Exceptionalism," *Social Sciences* 9, no. 12 (2020): 232. For a discussion of Western exceptionalism in the Danish context, see Mette Hjort, "The Epiphanic Moments of COVID-19: The Revelation of Painful National Truths," *Cultural Studies* 35, nos. 2–3 (2021): 505–513. For a discussion of exceptionalism in a number of countries, including Brazil, see Martha Lincoln, "Study the Role of Hubris in Nations' COVID-19 Response," *Nature* 585 (2020): 325.

opinion pieces by experts. Similar appeals were subsequently made in the leaders' approach to the development of COVID-19 vaccines and the implementation of vaccination programs. As we shall see, there has been an almost constant reference to a *leadership role* performed by the two nations and to the *love of freedom* in appeals to the British and American national character. 'Love of freedom' should be construed both as the freedom of individual citizens to choose for themselves and the importance of supporting the free market.

In order to better understand exceptionalism as manifested in beliefs about the superiority of one's nation and its capacity to control and manage significant threats, we turn to the literature on *unrealistic optimism* in section 4. People tend to be excessively optimistic about their own skills, talents, and virtues, and about their capacity to exercise control over their lives and avoid adverse events. Exceptionalism and unrealistic optimism support a positive self-image and foster feelings of belonging, helping people to manage negative emotions and sustaining their sense of agency. However, both exceptionalism and optimism give rise to epistemically irrational beliefs and may be conducive to taking excessive risks. In particular, it has been argued that they have cost lives in the context of the COVID-19 pandemic.[3]

Experts have argued that the idea that one's nation is better in some respect than other nations (e.g., better equipped to face a pandemic) has led to the delegitimization of medical advice and a refusal or reluctance to engage in international cooperation. This has prevented the UK and the US from learning from crisis management elsewhere and from responding in a timely and effective way to the challenges posed by the pandemic. As Cynthia Miller-Idriss said:

> In the case of COVID-19, populist nationalist leaders are thus more likely than other national leaders to reject scientists' advice, attack global organizations like WHO, promote scientifically unproven and potentially harmful treatments for COVID-19 and reject scientifically proven practices like wearing masks in public. Populist nationalist anti-elite and anti-science sentiments have undoubtedly led to higher COVID-19 infection and mortality rates as a result.[4]

3 See, for instance, Danny Haiphong, "The Great Unmasking: American Exceptionalism in the Age of COVID-19," *International Critical Thought* 10, no. 2 (2020): 200–213. For the effects of so-called national narcissism on the uptake of COVID-19 vaccines, see also Aleksandra Cislak et al., "National Narcissism and Support for Voluntary Vaccination Policy: The Mediating Role of Vaccination Conspiracy Beliefs," *Group Processes & Intergroup Relations* 24, no. 5 (2021): 701–719.

4 In Eric Taylor Woods et al., "COVID-19, Nationalism, and the Politics of Crisis: A Scholarly Exchange," *Nations and Nationalism* (2020): 1–19.

In section 5, we look at how exceptionalism has been used to justify poor responses to the pandemic by exploring the phenomenon of *confabulation*. Confabulations are explanations that are ill-grounded but offered sincerely, often to fill a knowledge gap. We suggest that confabulation was common when people were prompted to defend behavior that did not conform to the health and safety guidelines put in place to reduce the risk of infection. We notice how these confabulatory explanations appealed to the nationalistic themes of superiority and control and also to the ideals of personal autonomy and economic freedom. Similar to unrealistically optimistic beliefs, these confabulations may have had a significant psychological benefit by contributing to people's positive self-image; the very fact that people offered reasons for their behavior had some epistemic value, because it enabled their explanations to be discussed, reflected upon, and challenged. However, confabulatory explanations ultimately misrepresented potentially dangerous rule-breaking as an expression of love for freedom, contributing to people dismissing the effects of their behavior on their own safety and that of others, as well as on the containment of the virus.

To conclude, in section 6, we summarize the main points raised in the course of the discussion. Exceptionalism by itself is not a manifestation of irrationality: there may be good grounds to believe that one group (a nation in this case) is better equipped than other groups (nations) to respond to a threat. However, when claims of exceptionalism are grounded in nationalistic values and generalizations about national character, there are risks involved. These are the same risks individuals face when they unjustifiably inflate their conception of their own worth and expect not to suffer as much as others from setbacks. In the end, we focus on what the pandemic has taught us, considering what can be done differently in the future to avoid the risks of combining exceptionalism with nationalism.

2 British Exceptionalism

In February 2020, Boris Johnson, prime minister of the UK, argued that it was important that some governments in the world stood by freedom of exchange, contrasting this attitude with the "irrational" panic caused by the new coronavirus. In his speech, he indicated that the UK was such a country and compared it to a superhero ready to lead and save: "Humanity needs some government somewhere that is willing at least to make the case powerfully for freedom of exchange, some country ready to take off its Clark Kent spectacles and leap into the phone booth and emerge with its cloak flowing as the supercharged

champion, of the right of the populations of the earth to buy and sell freely among each other."[5]

Prior to the announcement of a lockdown in March 2020, the UK refused to follow the examples of China, Taiwan, and Korea, which had been imposing restrictions on their citizens to contain the spread of the virus. A writer for the *Guardian* newspaper observed that "rather than learning from other countries and following the WHO advice, which comes from experts with decades of experience in tackling outbreaks across the world, the UK has decided to follow its own path. This seems to accept that the virus is unstoppable and will probably become an annual, seasonal infection."[6]

On a number of occasions, Johnson explicitly suggested that British people loved freedom too much to tolerate restrictions on their movement, and so the lockdown measures adopted by other countries to contain the coronavirus could not be implemented. Commentators identified *love of freedom* as a thread in the distinct form of exceptionalism embodied by Johnson: "The myth of a unique and defining love of personal freedom as a badge of nationhood underpinned a profound reluctance to impose lifesaving restrictions on movement and social gatherings. Other people might put up with that sort of thing, but not the English. On the altar of this exceptionalism, lives have been sacrificed."[7] Interestingly, the UK did go into lockdown again after the first lockdown in March 2020, but the references to the country being special, unique, and second-to-none did not stop. In June 2020, Johnson invited people to enjoy the freedoms they had given up due to the COVID-19 restrictions and were now "rightly reacquiring," including going to shops and restaurants. Tory MP Gareth Johnson described people heading to the pub as "doing their patriotic best for Britain," asserting that going to the pub was a "great British institution" and was vital to getting the economy back on track.[8] The day when most COVID-19 restrictions were lifted in the UK, July 19, 2021, was named Freedom Day.

5 Boris Johnson, "PM speech in Greenwich: 3 February 2020," GOV.UK, https://www.gov.uk/government/speeches/pm-speech-in-greenwich-3-february-2020 (accessed February 1, 2021).

6 Devi Sridhar, "Britain Goes It Alone Over Coronavirus. We Can Only Hope the Gamble Pays Off," *The Guardian*, March 15, 2020, https://www.theguardian.com/commentisfree/2020/mar/15/britain-goes-it-alone-over-coronavirus-we-can-only-hope-the-gamble-pays-off.

7 Fintan O'Toole, "Coronavirus Has Exposed the Myth of British Exceptionalism," *The Guardian*, April 11, 2020, https://www.theguardian.com/commentisfree/2020/apr/11/coronavirus-exposed-myth-british-exceptionalism.

8 Neil Shaw, "Boris Johnson Says People Should Do 'Patriotic Best' and Go to Pub," *Wales Online*, June 23, 2020, https://www.walesonline.co.uk/news/uk-news/boris-johnson-says-people-should-18472969.

Examples of exceptionalism could also be observed in politicians' messages concerning the new vaccines against COVID-19, mixed with the need to persuade citizens that there were some advantages to Brexit at the end of the transition period, just before the UK exited the European Union. Health secretary Matthew Hancock claimed that the UK could approve the Pfizer/BioNTech vaccine for use without waiting for the European Medicines Agency. He said, "Because of Brexit, we've been able to make a decision to do this based on the UK regulator, a world-class regulator, and not go at the pace of the Europeans, who are moving a little bit more slowly."[9] Exultant statements about the UK being the first country in the world in which a COVID-19 vaccine was authorized and made available were accompanied by additional references to the potential beneficial effects on people's freedom and the national economy; the benefit was consistently cashed out in terms of "reclaiming our lives" and "getting our lives and livelihoods back."

3 American Exceptionalism

Similar themes of exceptionalism can be found in the handling of the virus in the US. Even before the virus hit, a belief in the superiority of the nation's capacity to tackle health threats could be seen in the decisions made to undervalue and cut funding toward projects based on pandemic prediction and preparation.[10] For instance, the Trump administration stopped funding a program aimed at providing alerts regarding potential pandemics a few months before COVID-19 infections started spreading in China.[11] The US did not act quickly in preparing to face the coronavirus, despite the fact that its severity was evident from the situation in other countries. Furthermore, in the midst of the pandemic, the decision was made to withdraw from the World Health Organization, undermining international cooperation in tracking the virus, producing vaccines, and protecting citizens. COVID-19 was described

9 Heather Stewart, Sarah Boseley, and Daniel Boffey, "Covid Vaccinations Will Begin Next Week, says Boris Johnson," *The Guardian*, December 2, 2020, https://www.theguardian.com/world/2020/dec/02/covid-vaccinations-will-begin-next-week-says-boris-johnson.

10 See Daniel Lippman, "DHS Wound Down Pandemic Models before Coronavirus Struck," *Politico*, March 24, 2020, https://www.politico.com/news/2020/03/24/dhs-pandemic-coronavirus-146884. See also Dan Diamond, "Inside America's 2-Decade Failure to Prepare for Coronavirus," *Politico*, April 11, 2020, https://www.politico.com/news/magazine/2020/04/11/america-two-decade-failure-prepare-coronavirus-179574.

11 Oliver Milman, "Trump Administration Cut Pandemic Early Warning Program in September," *The Guardian*, April 3, 2020, https://www.theguardian.com/world/2020/apr/03/trump-scrapped-pandemic-early-warning-program-system-before-coronavirus.

by President Donald Trump as being just "like the flu" and as something that one day, as a miracle, would disappear on its own.[12] Thus, Trump and the US appeared to be completely unprepared and incognizant of the seriousness of COVID-19 and the harm it would bring, optimistically assuming it would be easy to tackle.

Moreover, during the pandemic, Trump keenly emphasized the superiority of the US in responding to the virus, with no regard for the reality of how the country was actually faring in comparison to other nations. He claimed that the country was leading the way and even providing support to other countries: "I spoke with Angela Merkel today. I spoke with Prime Minister Abe of Japan. I spoke with many of the leaders over the last four or five days. And so many of them, almost all of them—I would say all of them; not everybody would want to admit it—but they all view us as the world leader, and they're following us."[13] Talking about ventilators, Trump claimed that the US was in a position to help other countries: "We have a very big stockpile right now. And we're building it bigger and we're helping a lot of other countries. Nigeria—we just sent a thousand. We have various—various countries: France, Spain. We have a lot going to Italy. We have a lot going to a different—probably 15, 18 countries. They're calling us. We had the capacity to do this; nobody else did."[14]

Trump's boasting reflects the general tendency for the US to present itself as the model for other countries to learn from, rather than being a country that looks to others. For instance, back in 2011, Mitt Romney described the US as having the strongest economy and the strongest military in the world. He said that "God did not create this country to be a nation of followers,"[15] and he too suggested that the US should lead the rest of the world. This idea that the

12 See Dan Mangan, "Trump Dismissed Coronavirus Pandemic Worry in January—Now Claims He Long Warned about It," *CNBC*, March 17, 2020, https://www.cnbc.com/2020/03/17/trump-dissed-coronavirus-pandemic-worry-now-claims-he-warned-about-it.html. See also Tommy Beer, "All The Times Trump Compared Covid-19 to the Flu, Even After He Knew Covid-19 Was Far More Deadly," *Forbes*, September 10, 2020, https://www.forbes.com/sites/tommybeer/2020/09/10/all-the-times-trump-compared-covid-19-to-the-flu-even-after-he-knew-covid-19-was-far-more-deadly/?sh=672450cdf9d2.

13 "Remarks by President Trump in Meeting with Republican Members of Congress," May 8, 2020, https://trumpwhitehouse.archives.gov/briefings-statements/remarks-president-trump-meeting-republican-members-congress/.

14 "Remarks by President Trump During Tour of Ford Rawsonville Components Plant," May 21, 2020, https://trumpwhitehouse.archives.gov/briefings-statements/remarks-president-trump-tour-ford-rawsonville-components-plant/.

15 Michael McGough, "God Made America, According to Romney," *Los Angeles Times*, October 7, 2011, https://opinion.latimes.com/opinionla/2011/10/god-made-america-according-to-romney.html.

US does not follow but leads, often emphasized by Trump, has led many commentators to view the nation's response to the virus as an act of "hubris."[16] The message is that US political leaders failed to grasp the seriousness of the problem and delayed taking action, and as a result their response was inadequate.

Trump often appealed to national identity and national values when downplaying the effects of the virus and the effectiveness of the required provisions. Rather than referring to expert advice or scientific research to justify his decisions, Trump appealed instead to American citizens' love of freedom. For instance, with regard to closing the country's borders, Trump said: "I don't think the [American] people are going to stand for it. This is a country that's meant to be open, not closed."[17] In promoting minimally invasive legislation for containing and tackling the virus, Trump appealed to the value of economic freedom.[18] The result was that the only area in which the US led the rest of the world was in the infection rates and number of deaths due to COVID-19.[19] Testing was plagued with setbacks and challenges, during which the virus continued to spread quickly,[20] and yet Trump claimed that the US tested more citizens than all other countries put together and furthermore utilized the "highest quality" test.[21] It was patently false that the US tested more than all other countries combined, and the accuracy and quality of the tests administered worldwide were comparable.[22] Trump also claimed that his administration had "taken the most aggressive action in modern history to prevent the spread of this illness in the United States."[23] This statement was obviously in tension with his initial dismissal of the pandemic, and again emphasized the superiority of the US in

16 See, for instance, Emily Tamkin, "Donald Trump's Hubris in the Face of COVID-19 Pandemic is Pure Americana," *New Statesman*, October 7, 2020, https://www.newstatesman.com/politics/2020/10/donald-trump-s-hubris-face-covid-19-pandemic-pure-americana. See also Uri Friedman, "Why America Resists Learning From Other Countries," *Atlantic*, May 14, 2020, https://www.theatlantic.com/politics/archive/2020/05/coronavirus-could-end-american-exceptionalism/611605/.

17 "Remarks by President Trump During Tour of Ford Rawsonville Components Plant."

18 Matthew P. Crayne and Kelsey E. Medeiros, "Making Sense of Crisis: Charismatic, Ideological, and Pragmatic Leadership in Response to COVID-19," *American Psychologist* 76, no. 3 (2020): 462–474, doi.org/10.1037/amp0000715.

19 See Haiphong, "The Great Unmasking."

20 Sara Murray, Nick Valencia, Jeremy Diamond, and Scott Glover, "How Coronavirus Testing Fumbles Squandered Valuable Time," *CNN*, April 20, 2020, https://edition.cnn.com/2020/04/20/politics/coronavirus-testing-trump-administration-response-invs/index.html.

21 Reality Check Team, "Coronavirus: President Trump's Testing Claims Fact-Checked," *BBC*, May 15, 2020, https://www.bbc.co.uk/news/world-us-canada-52493073.

22 Haiphong, "The Great Unmasking."

23 Mangan, "Trump Dismissed Coronavirus Pandemic Worry in January."

the face of mounting evidence that the national response and handling of the pandemic was inadequate and had catastrophic results.

The sense of superiority and the exceptionalism seen in the United States' response can be linked to historical nationalist values of liberty and freedom, just like the response by British political leaders. In both countries, the very same values have long been used to maintain systems that disadvantage many of their citizens, who as a result lack access to basic rights, including safe housing and healthcare. In the context of the threats posed by COVID-19, the emphasis on individual freedom led to an 'economy first' approach that caused immense harm and loss of human lives.[24]

4 Optimism about Health

Why do world leaders invoke exceptionalism in justifying their decisions, and why do many of their citizens follow them in believing that their country is an exception to the norm or superior to other countries? It is possible that the exceptionalism exhibited by political leaders is a means of propaganda, a show of strength and confidence, but we need a better understanding of the wider popularity and considerable persistence of exceptionalist attitudes and arguments across the rest of the population.[25] In this section, we want to draw some analogies between exceptionalism and unrealistic optimism, based on the fact that both tendencies can be used to justify denialist responses to a crisis.

It is common for agents to believe that something undesirable will not happen to them even when the undesirable event (such as a redundancy, serious illness, or divorce) is a frequent occurrence in their population or culture. This tendency is known as the *optimism bias* or *unrealistic optimism*.[26] Maybe the most cited example in the literature is that of divorce: in Western countries

24 Crayne and Medeiros write, "[T]he evidence suggests that countries led by more pragmatic leaders may have better long-term health outcomes from their response to COVID-19 than those with charismatic or ideological leaders. […] In contrast, attempts to understate the pandemic's seriousness and appeals for economic freedom that have been noted in Brazil are reflected also in the approach of leaders from the United States." "Making Sense of Crisis," 9.

25 This point is often raised in the literature on conspiracy theories: the motives of those who propagate the theories may not coincide with the motives of those who endorse the theories. See, e.g., Neil Levy, "Is Conspiracy Theorising Irrational?," *Social Epistemology Review and Reply Collective* 8, no. 10 (2019): 65–76, https://wp.me/p1Bfg0-4wW.

26 See Anneli Jefferson, Lisa Bortolotti, and Bojana Kuzmanovic, "What Is Unrealistic Optimism?," *Consciousness & Cognition* 50 (2017): 3–11.

divorce rates are extremely high (close to 50 percent), but when people are asked what they think their chance of getting a divorce is, they estimate it to be 1 or 2 percent.[27] They know that divorce is common, they just don't think it will happen to them.[28] Even 'experts,' such as divorce lawyers, are optimistic about their own marriages.

Another common tendency for agents is to believe that they are better than their peers in a number of domains including specific skills (such as driving or academic performance), general qualities (such as attractiveness and intelligence), and even moral character (such as generosity and altruism). This is known as the *illusion of superiority* or the *better-than-average effect*. There are several hypotheses about how the illusion emerges and is maintained: people tend to creatively interpret negative feedback and selectively remember successes while forgetting failures, and this can begin to explain why they may think of themselves as a better-than-average driver even after being involved in several car accidents, or why the overwhelming majority of academics regard their own work as better than average.

The optimism bias and illusion of superiority are often accompanied by a third, widespread bias, the *illusion of control*. Agents believe that they can control events that are independent of them and on which they exercise limited or no influence. Examples abound: in a betting situation, people overestimate their chances of winning when they themselves deal the cards or throw the dice; at pedestrian crossings, people assume that the green light is an effect of their pressing the button, whereas in many cases the lights follow a pre-established pattern.[29]

It is likely that the three biases interact with one another. Take Jei's belief that she is better than her peers in a number of domains and her conviction that she can control external events. Such attitudes will contribute to Jei's prediction that her future will not include adverse events. This is because

27 Tali Sharot, "The Optimism Bias," *Current Biology* 21, no. 23 (2011): R941–R945.

28 Some researchers have argued that this optimism is simply a manifestation of the person's commitment to the success of their relationship and to their romantic partners. What if the unrealistic optimism literature unveils aspirations instead of biased beliefs? See Owen Flanagan, "'Can Do' Attitudes: Some Positive Illusions Are Not Misbeliefs," *Behavioral and Brain Sciences* 32, no. 6 (2009): 519–520. This is an interesting perspective, and it can be shown to apply to some of the attitudes involved. However, from our point of view, what matters is not whether the attitude is a belief as such, but whether it has an impact on decision-making and drives action. As unrealistically optimistic attitudes do have pervasive effects on people's behaviors, their examination is an important contribution to a more thorough understanding of people's cognition and agency.

29 Stuart Vyse, *Believing in Magic: The Psychology of Superstition* (Oxford: Oxford University Press, 2014).

she believes that she can actively avoid adversities. The health domain is of particular relevance to our discussion: suppose that without any particular evidence to support her conviction, Jei comes to believe that she has a better immune system than average and that, if she adopts some safety behaviors, she can avoid contracting the virus altogether. As a result, in the context of a global pandemic, Jei confidently predicts that she will not catch COVID-19. This type of reasoning has been observed in other health contexts, including among breast cancer patients in remission who overestimate their capacity to control their health in the future and prevent the cancer from returning.[30]

How should we evaluate Jei's positive illusions? On the one hand, her prediction that she will not contract the virus is excessively optimistic, and her beliefs of superiority and control are epistemically irrational, in the sense that they are not robustly supported by the evidence at her disposal. On the other hand, her rosy prediction and optimistic beliefs have undeniable psychological benefits. Minimally, they can contribute to Jei's ability to successfully manage her own stress and anxiety in relation to the pandemic. More importantly, they can sustain her motivation to engage in those safety behaviors (washing her hands often, wearing a mask, maintaining social distance) that will protect herself and others from the virus.[31] If Jei has an unrealistically optimistic belief about the extent to which her safety behaviors can protect her from infection, she will comply with government advice enthusiastically and even encourage other people to do so. This will support her coping response more effectively than a fatalistic belief ("Nobody can avoid infection"), which may lead to disengagement.

But optimistically biased beliefs do not always result in coping effectively with threats and increased motivation in pursuing beneficial lifestyle changes; they could result in a denial of the threat. In that case, they may lift one's mood significantly in the short term, because the threat magically disappears, but they are conducive to risk-taking behavior with negative consequences for the individual and the community. If Jei convinces herself that she is immune to the virus, this may lead her to believe that adopting safety behaviors is not required. As a consequence, she might stop following the government's advice

30 Shelley E. Taylor, "Adjustment to Threatening Events: A Theory of Cognitive Adaptation," *American Psychologist* 38 (1983): 1161–1173.

31 Lisa Bortolotti, Magdalena Antrobus, and Ema Sullivan-Bissett, "The Epistemic Innocence of Optimistically Biased Beliefs," in *Reasoning: Essays on Theoretical and Practical Thinking*, ed. Magdalena Balcerak Jackson and Brendan Balcerak Jackson (Oxford: Oxford University Press, 2019), ch. 12.

to work from home when possible to reduce the spread of the virus and begin commuting to her office daily, putting herself and others at risk.

Attitudes toward the pandemic have already been explained by reference to unrealistic optimism:[32] researchers studying the attitudes of university students from three different countries (Poland, Iran, and Kazakhstan) at the time of the first and second waves of COVID-19 found that unrealistic optimism about the possibility of contracting the virus was a widespread and robust result.[33] Interestingly, the researchers found that medical professionals in Poland who had medical knowledge about COVID-19 and dealt with the consequences of the virus in their daily lives were not unrealistically optimistic. Further studies have confirmed the pervasiveness of unrealistic optimism in several European countries prior to the first wave,[34] when the risk of catching the virus was dramatically underestimated:

> As early as January 2020, renowned epidemiologists like Gabriel Leung or Marc Lipsitch had highlighted the threat of a global pandemic. [...] They announced that more than 40–70% of the world population could be infected within the end of the year. However, survey data collected in February 2020 during the early phases of the outbreak in France, Italy, the United Kingdom, and Switzerland showed that a large majority of citizens estimated their risk of catching the virus to be around 1%.[35]

The optimism bias may also incur 'systemic' costs. In her classic 2011 paper on the optimism bias, Tali Sharot argues convincingly that individual biases may be responsible for group behavior. In particular, she mentions the crisis of the financial market in 2008:

> The harmful influences of over-optimism likely extend to the collective behaviour of groups. For instance, the optimism bias has been named by

32 See Sinué Salgado and Dorthe Berntsen, "'It Won't Happen to Us'. Unrealistic Optimism Affects COVID-19 Risk Assessments and Attitudes Regarding Protective Behaviour," *Journal of Applied Research in Memory and Cognition* (2021), https://doi.org/10.1016/j.jarmac.2021.07.006.

33 Wojciech Kulesza et al., "We Are Infected with the New, Mutated Virus UO-COVID-19," *Archives of Medical Science* 17, no. 6 (2021): 1–10, doi:10.5114/aoms.2020.99592.

34 Hugo Bottemanne et al., "Does the Coronavirus Epidemic Take Advantage of Human Optimism Bias?," *Frontiers in Psychology* 11 (2020), https://www.frontiersin.org/article/10.3389/fpsyg.2020.02001.

35 Jocelyn Raude et al., "Are People Excessively Pessimistic about the Risk of Coronavirus Infection?," *PsyArXiv*, March 8, 2020, doi.org/10.31234/osf.io/364qj.

> several economists as one of the core causes of the financial downfall of 2008. Unrealistic expectation of individuals, financial analysts and government officials that the market would continue growing, despite evidence to the contrary, likely contributed to the collapse.[36]

Sharot considers several factors that can turn adaptive versions of individual optimism into causes of catastrophic events for groups or society at large. First, optimism bias is likely to increase with uncertainty (and both the oscillation of financial markets and the circumstances of a new global pandemic are to some extent difficult to predict, resulting in uncertainty). Second, the globalized nature of how information spreads, due to the internet and social media, means that an individual has the potential to communicate with (and influence) many more people. As Sharot puts it, "[I]ndividuals' biases that are inconsequential on their own can accumulate together to produce a large bubble."[37]

A similar transition from adaptive individual bias to negative systemic effects has been observed with respect to the problem of climate change inaction. When reviewing various factors that impact people's responses to climate problems, we find that people who score higher in optimism bias are also less concerned about the environment.[38] The case of climate change is especially interesting because of the time lag between action/response and outcomes. In some situations (such as the interval between taking an exam and receiving the results), people tend to 'shelve' their optimism when the 'moment of truth' comes, so they can prepare themselves for the outcome and avoid disappointment.[39] In the case of climate change, however, the moment of truth is too far into the future and the reality check does not apply.

This discussion has wide-ranging implications for attitudes toward the pandemic. The delay between action and outcomes is not as great as in the case of climate change inaction, because we can observe or infer the effects of a lockdown or a vaccination program on infection rates before too long. But there is still a significant time lag, and this means that any commitment to an action or response in the hope that it will have the desired effects requires something like a leap of faith. That is why excessive optimism can still lead people to

36 Sharot, "Optimism Bias."

37 Ibid., 944.

38 Sabine Pahl et al., "Perceptions of Time in Relation to Climate Change," *WIREs Climate Change* 5, no. 3 (2014): 375–388.

39 Kate Sweeny, Patrick Carroll, and James A. Shepperd, "Is Optimism Always Best? Future Outlooks and Preparedness," *Current Directions in Psychological Science* 15 (2006): 302–306.

ignore risks, underestimate the negative effects of inaction, and delay decisions that are beneficial in terms of pandemic outcomes but are economically or psychologically costly.

What are the analogies, then, between a country's exceptionalism and an agent's optimism? Nationalist values play the same role as self-enhancing beliefs in the unrealistic optimism phenomenon. Predictions about the future are unrealistically optimistic when the desired event requires a skill or talent that people mistakenly attribute to themselves (self-enhancement). If Sylvia thinks she is an excellent driver but is in fact deceived about her driving skills, her prediction that she will easily win the World Rally Championship is unrealistically optimistic. What about exceptionalism and the capacity to respond effectively to threats? When we believe that people and institutions from a given country are better equipped than others to deal with threats because of their history and culture, we may be right. There are cases in which a country's previous experience can confer advantages in dealing with threats: the fact that Taiwan's history involved a SARS outbreak, and that many Asian countries were already accustomed to mask-wearing, did put them in a better position when attempting to reduce COVID-19 infection rates.

But suppose that a country does not possess—or possesses to a lower degree than other countries—those virtues needed to respond effectively to a threat, even though such virtues are commonly associated with its national character. In that case, the prediction that, based on those virtues, the country's response to the threat will be effective—or more effective than that of other countries—is unrealistically optimistic. In the context of a pandemic, if a country falsely believes that it has a better health system than other countries, then the prediction that it will cope better with a health threat is unrealistically optimistic.

5 The Role of Exceptionalism in Confabulation

As we have seen, being unrealistically optimistic about one's chances of contracting COVID-19 and suffering significantly from it affects the actions and precautions one subsequently takes. On the national scale, the UK and the US were slow to react to the encroaching pandemic despite the fact that the damage and harm that it was causing in other European countries, such as Italy and Spain, was quite clear. At the individual level, many citizens did not update their beliefs about the likelihood of contracting COVID-19 despite the evidence that the virus was spreading quickly within their own countries, and they continued to underestimate the seriousness of the infection despite evidence

of significant health consequences. Such behavioral tendencies reflect the irrationality of unrealistic optimism; beliefs with desirable content are more resistant to change and less responsive to new contrary evidence than beliefs with undesirable content. Many went on with their daily lives despite the risks, continuing to meet with others, visit shops and gyms, and attend large gatherings without wearing masks or socially distancing. How can we explain reckless behavior despite the mounting evidence of the threats posed by COVID-19?

A leading factor in accounting for these behaviors is the prevalence of confabulation. When people explain their own behavior, they tend to confabulate: they come up with explanations and justifications of their choices and actions despite having a knowledge gap with regard to what actually influenced their behavior.[40] Crucially, they have no intention to deceive others when they do so and sincerely believe their own reports about what has driven their behavior. But they are motivated to give explanations that present themselves as rational decision-makers and preserve and emphasize their positive self-representations,[41] so they are less likely to consider explanations that may be more truthful but tell a less flattering story about themselves. In short, confabulations are ill-grounded explanations, so they often misrepresent the world and do not accurately capture the actual factors that are efficacious in determining people's behavior.

The influence of unrealistic optimism on people's behavior is a factor that people are unlikely to include in their explanations of why they acted and chose in the ways that they did, because they may not know how biases affect decision-making. Moreover, even if they did, acknowledging the influence of a bias would not cohere with their self-representations as reasonable agents. When people had to explain and justify behavior-flouting rules, they offered explanations such as, "If I get the virus, I get it. Even if I get ill, at least it's over and done with." This kind of thinking manifests unwarranted optimism: it disregards the possibility of suffering serious long-term symptoms of COVID-19, the possibility of death, and the possibility of passing the virus on to more vulnerable people. Other factors, such as being generally self-centered and irresponsible, were

40 On confabulation and consumer choice, see Richard E. Nisbett and Timothy D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84, no. 3 (1977): 231–259. On confabulation in moral reasoning, see Jonathan Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement," *Psychological Review* 108, no. 4 (2001): 814–834. On confabulation and justifying choices, see Lars Hall et al., "Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea," *Cognition* 117, no. 1 (2010): 54–61.

41 Ema Sullivan-Bissett, "Implicit Bias, Confabulation, and Epistemic Innocence," *Consciousness and Cognition* 33 (2015): 548–560.

probably efficacious in bringing this behavior about, at least in some instances, but again were not included in people's explanations of their own behavior. As we saw, the illusion of superiority applies to the domain of moral character as well, and people tend to describe themselves as more generous and altruistic than is warranted by the evidence, or more virtuous in general than the average person. This suggests that people may not have been fully aware of their own selfish motivations for action (or inaction). Even if they had become aware of their genuine motivation at some level, confabulation may have intervened, protecting and projecting positive self-images.

Exceptionalism may have been exactly what people needed to provide alternative explanations and justifications for their behavior—explanations and justifications that were ultimately ill-grounded but protective of a positive self-image. As we saw, the core of an exceptionalist belief regarding a nation is the idea that one's nation is an exception to the norm, where the anomaly is cashed out in terms of superiority (e.g., an increased capacity to control and respond effectively to threats). Exceptionalism was a recurrent theme in post-hoc justifications provided by people who were prompted to defend risky behavior or behavior that flouted the rules. Often, the exceptionalist justifications offered were based on the person's political and national identity and on their love for freedom. For example, anti-maskers in Florida defended their stance with references to religion, their "God-given right" to breathe, a rejection of communism, and an endorsement of democratic values, stating their case as "we the people."[42] Similarly, those who snubbed newly implemented regulations aimed at combatting COVID-19 often justified themselves along these lines: "This is a free country, the government can't mandate this extent of control."

Interestingly, these defenses often paralleled the themes of exceptionalism based on national identity emerging in Johnson's and Trump's speeches. As we saw in sections 2 and 3, both leaders emphasized the values of freedom, liberalism, and the free market as cornerstones of the national image in the UK and the US. The national identities constructed and portrayed by world leaders influence citizens' personal identities, in that the same themes are adopted in characterizing positive self-representations and are then featured in confabulatory explanations and justifications of rule-flouting. National identities and associated values become available as a source of personal identities and values, and thus are more likely to be featured in people's self-interpretations. For example, American citizens may draw on American values of economic freedom and personal liberties as God-given rights to defend their choice to attend

42 "Florida's Anti-Maskers Are Taking a Stand," *Yahoo News*, June 26, 2020, https://news.yahoo.com/floridas-anti-maskers-taking-stand-231500975.html.

gatherings despite the pandemic. They may believe that they are an exception *because they are American*, and Americans continue with 'business as usual' no matter what and do not learn from other countries but teach them instead.

The explanations and justifications people offer for personal behaviors that appeal to national values are epistemically problematic in several ways: (a) they tend to underestimate the severity of the threat; (b) they do not offer the whole picture of the motivation behind rule-flouting behavior, downplaying selfish interests, lack of responsibility, loss of patience, and recklessness; and (c) they may circumvent the need for people to better understand their own behavior.[43]

There are clear dangers and costs associated with confabulation that apply also in this context. On some accounts, confabulations consist of *representative reasons* given for agents' behavior, as opposed to the specific reasons that gave rise to it.[44] Representative reasons are reasons that are, by the standards of that society and culture, perceived as good and sensible reasons for a given behavior. In Richard Nisbett and Timothy Wilson's classic study on confabulation in consumer choice, research participants believed that they were involved in a customer survey. They were asked to choose among pairs of stockings and then they were asked the reason for their choice. According to the experimenters, most people chose the pair of stockings on their right-hand side due to a robust priming effect determined by the position of the items. However, research participants explained their choice of stockings by offering representative reasons for their choices ("I chose these stockings because they are softer/brighter"), that is, reasons that are widely perceived to be sensible ones in the context of consumer choice (such as preferred texture or color). In the experiment, the pairs of stockings to choose from did not differ in texture or color, yet this did not stop research participants from using representative reasons in the explanations of their choices.[45]

Similarly, culture and national identity could play a role in determining the representative reasons for certain behaviors in that culture or nation. In other words, the exceptionalism displayed by political leaders makes nationalistic reasons for actions widely available and salient, enabling their citizens

43 As we shall discuss shortly, the actual reasons why people act and choose the way they do may not always be available to them. However, motivated beliefs may still be relevant here, because among the reasons that are available for explaining and justifying behavior, people tend to select those that depict them in a positive light.

44 Nisbett and Wilson, "Telling More Than We Can Know."

45 For a discussion of the experiment and its significance in the philosophical literature, see Lisa Bortolotti, "Stranger than Fiction: Costs and Benefits of Everyday Confabulation," *Review of Philosophy and Psychology* 9, no. 2 (2018): 227–249.

to deploy them as representative reasons for their own behavior. In the case of controversial behaviors that people are particularly keen to defend from challenges, these 'ready-made' and highly regarded justificatory reasons are likely to be used. Confabulations that make use of reasons that are consistent with one's national identity and embody national values are less likely to be scrutinized and criticized by peers, with the effect that even dangerous and irresponsible behavior is seemingly condoned and can go unchallenged for longer. In contexts where national ideals are overvalued and may not always receive the scrutiny they deserve, they become a powerful source for narratives used to justify behavior. This prevents the potentially hypocritical mismatch between values and behavior from becoming apparent: the refusal to engage in safety behaviors is seen as a consequence of valuing individual freedom for oneself and one's fellow citizens, and not as an open disregard for one's own and one's fellow citizens' well-being in the face of an increased risk of infection.

Confabulatory explanations that appeal to national character and national values, or to personal values that are consistent with national character and national values, also justify behavior in a way that fends off potential allegations of unrealistic optimism. It is not excessively optimistic to believe that British and American citizens will not suffer from COVID-19 as much as citizens of other countries because exceptionalism says that the UK and the US are better placed to face the pandemic. Similarly, it is not unreasonable to avoid lockdowns and refuse to wear masks because exceptionalism says that in the UK and the US freedom is paramount and cannot be compromised—"it is in the country's DNA," and no threat is serious enough to justify a limitation of that freedom.

When people interpret their behavior in a way that is consistent with a positive self-image but distorts reality, confabulation incurs various epistemic costs. Scientific evidence is dismissed, the world is misrepresented, psychological factors actually driving behavior are ignored, and potential hypocrisy is masked. In section 4 we compared situations in which unrealistic optimism is fruitful and situations in which it is costly. In the case of COVID-19, the seriousness of the threat was minimized in a way that led some countries to remain underprepared and prompted some political leaders to reject expert advice on national policy. At the individual level, effects included a mistrust of scientists and a refusal to comply with health and safety recommendations. In sum, reckless and harmful behavior continued because people justified it to themselves in line with highly valued personal and national ideals, putting lives at risk and weakening attempts to control the spread of the virus.

So, confabulation means that some people found ready-made justifications for their irresponsible behavior in the exceptionalist rhetoric of their leaders.

However, the picture need not be so bleak. We should also consider the positive aspects of confabulating. Just as unrealistic optimism can lead either to effectively coping with a threat or to denying it altogether, confabulation can have positive as well as negative effects. Why do people confabulate? Confabulation is driven by a fundamental need people have to explain their behavior to themselves and to others, and to share the reasons for their actions and choices with their peers in a social interaction.[46] However, people cannot explain their behavior on the basis of causal mechanisms that are unknown to them, and cannot share information that is not available. Thus, in some circumstances, an accurate, well-grounded story about why people think, act, and choose the way they do is completely or partially unavailable to them. If agents' behavior is due to implicit biases or environmental cues of which they are unaware, those biases and cues cannot appear in the explanations they offer for their behavior. This means that, sometimes, the alternative to confabulating the reasons for one's behavior is not to provide an accurate explanation for it, because, as we have seen, agents may not have access to the factors influencing or determining their behavior.[47] Rather, the alternative is to provide no explanation at all, or to reply "I don't know" to questions about why one acts or chooses the way one does. This response would cut short the process of seeking understanding and inquiring about reasons or causes with others. Although the widespread nature of confabulation reveals a lack of insight into exactly why people behave the way they do, it also shows that people have a constant motivation to gain that insight. If they had no interest in understanding their behavior, this would be more disturbing than confabulating, especially when behavior can have such critical and far-reaching consequences.

It is epistemically valuable to be motivated to find explanations for one's behavior and to share those explanations in social interactions. If people had no interest in understanding, explaining, and justifying their behavior, then they would feel no pressure to give reasons for rejecting masks, social distancing, lockdown restrictions, and so on, and their reasons could not be available for personal reflection and for external scrutiny. Only when explanations are offered and shared can they be taken apart and assessed, and ultimately endorsed or rejected. When people engage in reason-giving practices, their

46 Sophie Stammers, "Confabulation, Explanation, and the Pursuit of Resonant Meaning," *Topoi* 39, no. 1 (2020): 177–187. For a discussion of the social ends of cognitive functions, see Carolyn Parkinson and Thalia Wheatley, "The Repurposed Social Brain," *Trends in Cognitive Sciences* 19, no. 3 (2015): 133–141.

47 Lisa Bortolotti, *The Epistemic Innocence of Irrational Beliefs* (Oxford: Oxford University Press, 2020), ch. 3.

attention can be drawn to the coherence and overall plausibility of the reasons they offer. For example, their attention can be drawn to the fact that behavior endangering their own and other people's lives may not actually embody the values of personal freedom and autonomy.

6 Lessons from the Pandemic

In this paper, we discussed the role of exceptionalism in justifying slow responses by some countries to the COVID-19 pandemic in 2020 and moved on to draw some analogies between exceptionalism as a feature of nations and unrealistic optimism as a widespread bias in human cognition. We remarked how optimism has both costs and benefits for people who make excessively positive self-assessments and predictions about their own future, and how it turns bad when it leads people to deny the seriousness of a threat rather than motivating them to mount an effective response to it.

Exceptionalism has not just served as an alibi used by political leaders to justify ineffective policies for the containment of COVID-19, but has also influenced the way individual citizens justify their rule-flouting behavior when refusing to conform to safety recommendations and guidelines. When asked why they were not using masks or staying at home, people often referred to the same ideals of superiority and freedom that their political leaders had endorsed. It is a good thing that people have a strong need to interrogate themselves about why they behave the way they do and to share their explanations with others. However, some of the factors responsible for human behavior are unavailable, either because people are unaware of their own biases or because they are motivated to ignore such biases. In those instances, people may explain their behavior in ways that are ill-grounded and overly flattering.

Given our discussion of the role of exceptionalism, the costs and benefits of optimism, and the inevitability of confabulation, can anything be done to avoid in the future the mistakes that were made in tackling the first wave of the COVID-19 pandemic? The reactions to the health emergency from nations and the public alike have highlighted how national values, unrealistic optimism, and confabulation have pervasive effects on national policy and individual behavior. At the personal level, the likeliness that reason-giving will aid self-understanding in the long run rests on agents' humility about their self-regulatory practices. It is important for agents to recognize that they are fallible and might not always be living up to their own ideals or to the political and national values they are committed to. Having such humility will also bring a healthier and more open attitude toward expert advice and toward the

feedback agents receive from others on their own behavior. Indeed, the need to believe that one has a special claim to knowledge about a certain event and the lack of epistemic humility have been identified by cognitive scientists as factors contributing to the adoption and spread of conspiracy theories and the general mistrust of scientific expertise.[48]

Further, it is possible to turn agents' hardwired optimism into a force for good, if realistic predictions from experts are taken seriously and agents are empowered to think that they can control *something*, even in scenarios of great uncertainty. For instance, they can control their own behavior and make their own contributions to efforts to contain and overcome threats. Angela Merkel, the German chancellor, was praised for doing exactly this.[49] In her communications with the public, she regularly referred to scientific findings while also allowing for some hope and optimism. For example, she stated that "It is true that the latest numbers . . . as high as they are, very cautiously give a bit of hope. However, it is definitely too soon to recognize a definite trend, and it is way too soon to start loosening any of the strict rules we have imposed on ourselves."[50] Cautious optimism can bolster preparations prior to facing a threat, rather than turning into a denial of its seriousness.

Merkel was also praised for taking heed of other countries and their successes in tackling the virus.[51] In particular, she recognized the history and prior knowledge of South Korea, given that the country had tackled a different coronavirus five years before. The value of international cooperation has proven critical in addressing the pandemic. At the level of national values and national character, it would be beneficial to have a more constructive set of values to identify with—values that emphasize the importance of cooperation and reserve a role for scientific expertise. This attitude has already been explicitly advocated in discussions about the rollout of the vaccine: "Global cooperation on vaccine allocation would be the most efficient way to disrupt the spread of the virus. It would also spur economies, avoid supply chain disruptions, and prevent unnecessary geopolitical conflict. Yet if all other

48 Roland Imhoff and Pia K. Lamberty, "Too Special to Be Duped: Need for Uniqueness Motivates Conspiracy Beliefs," *European Journal of Social Psychology* 47, no. 6 (2017): 724–734, https://doi.org/10.1002/ejsp.2265.

49 Crayne and Medeiros, "Making Sense of Crisis."

50 "Angela Merkel Sees 'A Bit of Hope,' But Keeps Coronavirus Lockdown in Place," *DW News*, April 3, 2020, https://www.dw.com/en/angela-merkel-sees-bit-of-hope-but-keeps-coronavirus-lockdown-in-place/a-53010223.

51 David Rising, "Germany Praised for Handling of Covid-19," *Yahoo News*, April 23, 2020, https://news.yahoo.com/germany-praised-handling-covid-19-044913509.html.

vaccine-manufacturing countries are being nationalists, no one will have an incentive to buck the trend."[52]

Optimism with regard to oneself and exceptionalism with regard to one's nation do not inspire successful responses to a crisis unless the values attributed to oneself and one's nation include epistemic humility and cooperation. Without the uncritical belief that one is superior to others and an exception to the norm, natural optimism can be more conducive to effectively coping with difficulties and embracing safety behaviors, and the inevitable confabulations are less likely to disguise reckless rule-breaking or unscrupulous self-interest as a defense of freedom.

Acknowledgments

The authors would like to thank the members of the Work in Progress group at the University of Birmingham, and in particular Al Wilson, for comments on a previous version of the paper. Also, Kathleen Murphy-Hollies is grateful to the audience of the Cognitive Distortions and Democratic Failures workshop at Università del Piemonte Centrale, Italy, for excellent feedback on her presentation of the paper.

52 Thomas J. Bollyky and Chad P. Bown, "The Tragedy of Vaccine Nationalism," *Foreign Affairs*, September 2020, https://www.foreignaffairs.com/articles/united-states/2020-07-27/vaccine-nationalism-pandemic.

Are Conspiracy Theorists Confabulating?

Key words: Confabulation; conspiracy theories; pathology; rationalisation; justification

Abstract:

In this paper, I argue that conspiracy theorists, in making conspiracist claims, should be understood to be confabulating, and I outline the ways in which this is helpful and illuminating. I show how three features in particular of conspiracist claims are better understood when the agent behind them is understood to be confabulating. These are that conspiracist claims (i) are always 'chopping and changing', (ii) often draw on further conspiracy, and (iii) can even endorse contradictory positions. Understanding agents to be confabulating illuminates the everyday cognitive architecture and motivations which go into making and endorsing conspiracist claims.

Given the bizarreness of many conspiracist claims, the lack of evidence for them, and how irrational they can be, it is common to think that they signal pathology on the part of the agent. However, I argue that when conspiracist claims are understood as confabulations, we have an explanation for them which captures those features and yet doesn't need to posit pathology despite, as I will explain, it still being possible for conspiracist claims to appear alongside pathology. They are not, however, indicators of it in the ways commonly thought.

Introduction:

In this paper, I argue that conspiracy theorists, in making conspiracist claims, should be understood to be confabulating, and I outline the ways in which this is helpful and illuminating. As the outputs of confabulating, I refer to conspiracy claims as 'confabulations', but my claim is that conspiracy theorists are *confabulating*, rather than that conspiracist claims themselves are confabulations no matter the context of uttering them. Given the bizarreness of many conspiracist claims, the lack of evidence for them, and how irrational they can be, it is common to think that they signal pathology on the part of the agent. However, I argue that when conspiracist claims are understood as confabulations, we have an explanation for them which captures those features and yet doesn't need to posit pathology despite, as I will explain, it still being possible for conspiracist claims to appear alongside pathology. They are not, however, indicators of it in the ways commonly thought.

By conspiracist claims, I mean expressions of belief[1] in claims which are bizarre and outlandish, and part of a conspiracy. For example, a belief that vaccines contain small computer chips which

[1] I assume a doxastic view of conspiracy theory ideation, such that conspiracies theorists do *believe* those conspiracy theories. See Ichino & Räikkä (2020) for an opposing view.

track people. Or, that the Earth is flat and NASA deceives us. Or, that a mysterious figure, 'Q', communicates information to certain people about a child sex trafficking ring which comprises many democrat politicians. These sorts of claims are, for most of us, recognised as bizarre, unsupported, and even harmful. However, confabulations too can make very poorly supported claims about the world, and looking more carefully at processes involved in confabulation can, I suggest, also make sense of how and why these conspiracist claims are made.

In section 1, I outline certain features of confabulations which are discussed in the literature and settle on some key ones, which can be used as indicators of confabulation. Then, in section 2, I show how these same features can be seen in, and make sense of, how individuals come to believe conspiracies and make conspiracist claims. In section 3, I explain how understanding conspiracist claims as confabulations captures features of those claims which often prompt us to think that they signal or manifest pathology on the part of the agent. Understanding conspiracist claims as confabulations uncovers the everyday psychological architecture and motivations behind forming these very bizarre beliefs, without needing to posit pathology. However, given that confabulation can also be a part of pathology, it doesn't completely rule out the possibility of pathology being present in those with conspiracist beliefs. Yet, this would merely be a case of the contents of these beliefs (which concern conspiracy), being found within a framework of clinical confabulation. The features which we thought were reasons to think that pathology is present, then, such as the bizarreness of the claims, are not the signals of pathology.

1. What is confabulation?

An agent can be said to have confabulated when they offer post-hoc, ill-grounded reasons for some past choice or behaviour of theirs. The reasons given are not based on the relevant evidence and so are ill-grounded in two senses; they don't faithfully track the actual causes of one's behaviour, and they are also unsupported by evidence. This means that they are usually false claims, barring some circumstances where they are accurate but the claim was still not made *on the basis* of the evidence available. Here is what is taken to be a paradigmatic case of confabulation to show this.

In an experiment by Nisbett and Wilson (1977)[2], participants were asked to choose the best pair of nylon stockings from a range presented to them, under the guise of a 'consumer survey'. They were then asked to give the reasons for their choice. In their explanations, participants would often say that their chosen stockings had superior colour or fabric. However, in reality they were

[2] I am wary of the possibility of this specific study failing to replicate. In any case, other evidence of people's tendency to rationalise post-hoc abounds.

all identical. There was in fact a pronounced bias for choosing stockings on the right-hand side, but this factor was not mentioned by any of the participants. Explanations which referred to the stocking's better colour or fabric therefore, were ill-grounded because they were firstly not based on the evidence of the stockings in front of them (they were all identical and so statements about a superior colour or fabric are false), and secondly they did not capture factors which were causally efficient in bringing them to choose as they did (that the stockings chosen were on the right hand side).

There are a few other important features of confabulation to note here. The first is that there is no intention to deceive. The participants are putting forward an account of their behaviour which they completely endorse and are confident in. Studies show the people are no less confident in their confabulations than in more accurate accounts of their behaviour (Johansson 2005; 2006). Another is that confabulation fills a cognitive gap (Hirstein 2005; Sullivan-Bissett 2015, 552). That is to say, the participants do not have cognitive or introspective access to the real cause of their behaviour; in this case, the influence of the position of the stockings. Why people tend to fill this cognitive gap with confabulation when they are asked to explain their choice, rather than notice their lack of answer, is the key mystery of confabulation (Bortolotti 2018, 237). However, a number of possible motivations go some way to answering this question, such as a desire to have a causal understanding of events (Coltheart 2017), to avoid the embarrassment of being dumbfounded and without an answer about oneself (Haidt 2001), to construct and enhance a sense of self and protect positive self-concepts (Bortolotti 2018; Sullivan-Bissett 2015), to place events into a meaningful narrative (Örulv and Hydén, 2006), to signal one's rationality and trustworthiness (Bergamaschi Ganapini 2020), and finally to facilitate social connection and share resonant themes with others (Bortolotti 2018, 242; Stammers 2020). I will return to some of these motivations in the context of making conspiracist claims.

Now I will discuss two features which are more contentious: do confabulations consist of causal reasons or motivating reasons? And, are confabulations explanations or justifications?

Causal reasons are external reasons *that* something happened. For example, the reason *that* the bridge collapsed is that the lorry was too heavy. Motivating reasons are an agent's own reasons *for* having done something. In other words, the reason for which she is taking herself to have acted. For example, Sarah's reason *for* shouting at her friend was that she was offended at what he said. An important distinction is that motivating reasons are the reasons which the agent takes herself to have acted on. Causal reasons for an agent's behaviour are often be overlooked and motivating reasons are favoured. It is easy to imagine, for example, that a causal reason for Sarah's outburst at her friend is that she is tired and stressed, and therefore easily

irritated. But these are less likely to feature in her explanation; instead, she gives motivating reasons. She states that her reason *for* shouting is that she was offended by her friend.

We tend to see that individuals explain their own behaviours with motivating reasons but explain other people's behaviour with causal reasons (Keeling 2018, 1217-1218; see Malle et al 2007 for discussion of why). In-fitting with this, confabulations tend to consist of motivating reasons. The positioning of the stockings is a causal reason for the participants' choices which they do not capture in their explanations, instead giving their own motivating reasons. Their reason *for* choosing the stocking was that it was a nicer colour than the rest.

These kinds of considerations have led some to criticise confabulation studies for conflating causal and motivating reasons (Sandis 2015), and to argue that it is hasty to conclude on the basis of them that we have very limited awareness and knowledge of the causes of our own behaviour (Jongepier and Cassam 2020). They suggest that really, it is not particularly strange that participants offer motivating reasons rather than causal reasons for their behaviour. And furthermore, they are perhaps more accurate in doing so. After all, in a sense, participants are surely right to say that being brighter or softer was *their* reason for having chosen that particular stocking.

However, being a motivating reason doesn't mean that the explanation isn't, or can't be, ill-grounded. The fact of the matter is still that the stockings are all the same colour, and even this motivating reason is unlikely to have actually been the motivator of their action. This can be seen most explicitly in choice blindness studies. In these studies, participants express a preference. This might be for a particular type of jam (Hall et all, 2010), or for a particular political opinion (Strandberg et al, 2018). Then, the investigator asks the participants to give a reason for their choice, but they recite the wrong choice back to them. So, participants are asked to give reasons for a choice that they did not actually make. Despite this, participants will often do so. Even if these are motivating reasons, they cannot actually have been their motivating reasons for the jam or political opinion chosen, because it is not the jam or political opinion that that they chose!

Not only this, but ethical concerns can mean that we are not always so satisfied with motivating reasons truly being the reason which the agent takes herself to be acting under, especially when there are certain important causal reasons in play. For example, people might confabulate motivating reasons for a choice which actually stemmed from the operation of implicit bias (a causal reason) (Sullivan-Bissett 2015, 550). In these cases, we are less inclined to be satisfied with the claim that someone's race or sex wasn't, technically, the agent's reason *for* displaying prejudiced behaviour, and that their motivating reason was indeed that they found the individual in question underqualified or incompetent. Here we see the importance of

acknowledging and aiming to capture some causal reasons for behaviour, even if this takes work. Therefore, I acknowledge that confabulations tend to express motivating reasons, but this doesn't allay many worries which confabulations bring, along epistemic or moral lines (examples of this are discussed in Bortolotti & Murphy-Hollies (2022), and Murphy-Hollies (2022)).

So far, I have referred to confabulations as 'explanations' of behaviour. However, some suggest they are better understood as 'justifications' (Bergamaschi Ganapini 2020). That is, the confabulating agent's focus is on normatively justifying their decision rather than giving an accurate explanation of the cause. Accepting that confabulations tend to be motivating reasons appears to sway us in the direction of seeing confabulations as justifications (this would be why motivating reasons are favoured by agents – because they provide better justification), but I don't think that this is the whole story.

Bergamaschi Ganapini (2020) argues that the participants in Nisbett and Wilson's study are unsatisfied with the explanation that their choice was due to the stockings being on the right-hand side because it doesn't normatively *justify* that choice for them. This is despite it being a perfectly acceptable and plausible causal explanation. Therefore, participants seek these kinds of justificatory accounts of their own behaviour, rather than cold, causal, explanatory accounts (2020, 193).

However, I think that individuals can, sometimes, be similarly comforted by causal explanations, and be drawn to them to the extent that they may sometimes feature in confabulations. For example, people are commonly drawn to explanations of their behaviour which draw on horoscopes and one's star sign. It doesn't make sense to say "my reason *for* being bossy is that I'm a Leo", and so horoscopes don't give rise to motivating reasons, but rather causal ones. People seem to recognise one's star sign as a reason *that* they are bossy. And this nevertheless seems to fulfil these kinds of normative justificatory requirements which they seek to meet (it is easy to imagine that another, less flattering factor could more accurately be the cause of someone's being bossy). Other examples could be seeking psycho-analytic style explanations for one's behaviour, seeking psychiatric diagnoses, or seeking neuroscientific and evolutionary explanations of behaviour. Despite these things being explanations with causal reasons, they can and do justify behaviours in the same way as motivating reasons. Normativity is at play in both. We want to justify our behaviour but we also want to adequately explain it. After all, we do see that individuals accept and respond to provocations by others when pressed on the *explanatory* inadequacy of their responses. They elaborate further, even if the original answer was a perfectly good justification. We will also see that agents elaborate further if good

explanations are nevertheless not good enough justifications for their behaviour, in their eyes. So, I propose that confabulations are both justifications and explanations[3].

To take stock of this section, I have said that confabulations are ill-grounded, motivating reasons for behaviour which are given post-hoc, and which aim to explain but importantly to also justify those choices. In the next section, I move on to look at features of conspiracist claims, and see if they can be understood as confabulations. That is: are they ill-grounded accounts of why the individual holds certain views, which seek to explain and justify them with motivating reasons?

2. Features of conspiracist claims

I draw on Quassim Cassam's (2019) plausible preliminary outline of what a conspiracy theory is. He distinguishes 'conspiracy theories' from 'Conspiracy Theories'. The former are theories of conspiracies which are reasonable to hold, such as Guy Fawkes' conspiracy to blow up parliament. These kinds of claims are not the kind I am interested in in this paper. Instead, I am interested in claims which relate to Conspiracy Theories, as Cassam refers to them. Conspiracies of this latter type are far less reasonable and more outlandish. As Cassam notes, they are speculative, contrarian, esoteric and amateurish (2019, 28). They are based on conjecture rather than solid evidence, favouring superfluous explanations over simpler ones, which go against general consensus about the events in question. It is also common to see that Conspiracy Theories often have, built into them, defences against any possible contrary evidence or considerations. There will always be answer for why some counterevidence exists, such as by positing that large-scale deception is at play by organisations who are 'in on' the conspiracy. This is sometimes described as being 'self-sealing' (Lewandowsky & Cook 2020), meaning that any counterevidence to the theory becomes re-interpreted as evidence *for* that theory.

Here's an example of a conspiracy theorist espousing conspiracist claims. Although the example is constructed, I aim to incorporate features of typical conspiracy theorists which have been found in empirical studies. Throughout this paper, I will note features of these claims which can be better understood if they are taken to be confabulatory.

Robin thinks that the new vaccine for covid-19 isn't safe and refuses to get it herself. She thinks that the procedures for manufacturing a safe vaccine have been rushed and therefore not completed properly. She thinks that pharmaceutical companies are making lots of money from giving people unsafe vaccines and are also taking money from billionaires to put tracking chips in the vaccine, and that's why authorities are so keen for everyone to get it. The coronavirus hasn't

[3] Stammers (2020) has a similar account, arguing that confabulations are explanations that are imbued with themes which resonate with us. Then, we share and communicate these themes and values with others.

been around long enough for scientists to be able to make a vaccine, and so the virus itself is just the flu but is exaggerated by politicians who want to keep people scared and at home. She also believes that given that it first came from China, it was maliciously released from there as a biological weapon. Robin's friends challenge her on some of these beliefs, telling her that the vaccine is safe and even point her to scientific papers which show that the vaccine has been subject to just as vigorous testing as all other vaccines. But Robin doesn't trust that those scientists' papers are legitimate, and thinks that they have ulterior motives in trying to show that the vaccine is safe. She believes that the science also shows that vaccines lower fertility, alter people's DNA, and can give them autism. The scientific community hides this research from the public.

Robin does not aim to deceive the people around her, and she completely endorses not only the conspiracist claims themselves as true, but also that their being true about the world is the reason that she endorses them (in other words, that the belief is based on evidence). I assume that agents often undergo the *transparency method* (Evans 1982) when giving reasons for their adoption of conspiracist claims. That is, when asking them why they hold those beliefs, they 'look through' their mental states to name things about the world. I take this not to be unusual[4], and how self-explanation often works. The participants in Nisbett and Wilson's (1977) study, for instance, 'look through' themselves to single out features of the world (softness of the stockings) in explaining their choice. Because of this process, explanations about oneself often double-up as explanations of the world as well. (Though not always. Other participants could have said 'I personally like soft socks'). Robin's reasons for her conspiracist beliefs about the covid vaccine not being safe, i.e., that there hasn't been adequate testing and that politicians and billionaires are deceiving the public for their own benefit, are ill-grounded. Not only are they false statements about the world, but as I will go on to demonstrate, they are not the whole story of how and why Robin has come to endorse these conspiracist claims.

2.1. Feature one: 'Chopping and changing' and elaboration.

The first feature here in the example of Robin which I will discuss is that the conspiracist claims are always 'chopping and changing'. Whenever an individual is pressed on why they believe some conspiracist claim, they tend to eagerly elaborate and defend their views. Given the specialist knowledge required to refute such claims, and the numbers of theories out there to contend with, interacting with conspiracy theorists has been compared to playing 'whack-a-mole' (Cassam 2019, 102). In Robin's case, in response to pressure put on her beliefs about

[4] Common enough and discussed by enough accounts of self-knowledge that my arguments in this paper do not, I hope, depend on any specific theory of self-knowledge.

vaccines from others and from counterevidence presented to her, she makes further claims that powerful organisations are publishing pushing phony research which shows that the vaccine is safe, while concealing good research about the vaccine's harmful side-effects.

Robin is provoked by others, and in response to this social provocation, she elaborates on her views in a way which protects them. Confabulation too is something that is brought on following provocation by others in this manner (Sullivan-Bissett 2015, 552; Hirstein 2005, 21). Agents may not have thought much at all about some behaviour of theirs before being asked, but once being asked, they do not want to not have an answer.

Crucially, provocation does not always have to be as explicit as being asked, outright. In fact, many conspiracy theorists may come to believe in conspiracies in quite a passive way; through being in an echo chamber, for example, and it is simply expected that everyone in that echo chamber believes, for example, that climate change is a hoax. However, provocation is still subtly at play. If you were not to go along with the status quo in these environments, there would be social penalties. This goes to show that subtle social provocation still takes place. In any culture, there are expectations that individuals are able to explain the choices and stances which they take on what are perceived to be important contemporary issues. (They may of course not participate and not make any choices or take a stance, but then they are not conspiracy theorists and so are not the concern of this paper). But for sure, no-one in an echo chamber is going to say that they believe climate change is a hoax because they are in an echo chamber with strong social pressures to believe as such, and with limited access to other sources. In sum, the initial provocations and first embracing of conspiracist claims can be very subtle and embedded quite implicitly within social interactions. Even when scrolling through social media alone, and coming across many nefarious claims about the science of climate change, an individual may start to feel that social pressure to take a stance on those issues.

In Robin's case we also see secondary confabulation, and this is more likely to be more explicit. This is when, in response to additional questions and pressure from others, people offer further confabulations to justify their primary, or initial confabulation (Bortolotti & Cox 2009, 954). Secondary confabulations are often ways to explain away inconsistencies. An example of this is discussed by Moscovitch (1995), in the case of a man with severe memory impairment. Despite having been married to his wife for 30 years and having grown children, due to his memory impairment he believes that they have only been married for four months. When others press this belief by pointing out that he has children who are adults, he offers the secondary confabulation that the children are adopted (which was not true). In Robin's case, she explains away the inconsistency of scientific research showing that the vaccines are safe by undermining that research as fake. Other scientific research, which highlights harmful side-effects and is

concealed, is not fake. So, further provocation and secondary confabulation is likely to be brought on more explicitly, by others who more clearly intend to put pressure on those ideas.

2.2. Feature two: Drawing on further conspiracy.

This quickly brings us to the second feature of conspiracies which I will discuss. Robin draws on further conspiracies in elaborating on and protecting her first conspiracist claims. It is common to see that people who believe in one conspiracy theory, tend to believe other conspiracy theories too. For example, research has shown that people who believed a conspiracy theory about the 9/11 U.S. terrorist attacks were more likely to believe other, unrelated conspiracy theories (Swami, Chamorro-Premuzic, & Furnham, 2010; Swami & Coles, 2010). The same has been found in those who endorsed a conspiracy theory about the 7/7 London terrorist attacks, and the strongest predictor of belief in even a fictitious conspiracy theory was belief in real-world conspiracy theories (Swami et al, 2011). These findings, showing that endorsing one conspiracy theory predicts and increases the chances of endorsing other conspiracy theories, have been taken to support Goertzel's (1994) suggestion that conspiracy theorists have 'monological belief systems'. This means that they develop a whole worldview in which nearly anything can be explained by reference to conspiracy and wide-scale deception. However, the structure of this monological thinking ought to be understood not so much as a web of lots of conspiracy-based explanations mutually supporting each other, but rather as a *style* of belief, or a central interest of the person to which everything gets linked (Klein, Clutton & Polito, 2018).

Understanding conspiracist claims as confabulations greater illuminates the nature of this monological thinking, and also situates it within more common and recognisable human psychology. I have mentioned some of the motivations involved in explaining why people confabulate at all, and why confabulations have the specific content that they do (Sullivan-Bissett 2015, 552). People strongly prefer to be able to give an answer and therefore confabulate one (remember that confabulations fill a cognitive gap, and so giving an accurate answer isn't an available option), rather than have no answer and endure the embarrassment and discomfort of being dumbfounded (Haidt, 2001). Then, the content of confabulations protect and enhance one's sense of self as a coherent agent, and one's own positive self concepts (Bortolotti 2018; Sullivan-Bissett 2015, 552). In general, confabulations are contributions to an ongoing narrative we have of ourselves, which make our own values and characters more explicit to ourselves and others. Although the claims made are ill-grounded and justified in a post-hoc way, they are united in conveying and demonstrating higher-order beliefs about the self, as being coherent and cognisant. This is the central interest which the content of confabulations, despite varying in specific details, always speaks to; being a coherent agent who knows herself and the world.

When it comes to conspiracy theories also, not only is it likely that similar motivations are in play as those suggested to underlie confabulation (particularly, to have a causal understanding of events (Coltheart 2017), to signal one's rationality (Bergamaschi Ganapini 2020) and for social engagement and connection (Bortolotti 2018, 242-243; Stammers 2020; Haidt 2001)), but they have the same structure. Both confabulations and conspiracist claims are made in aid of supporting higher-order beliefs, about the coherency of the self and of the world. Namely: I am a coherent, integrated and competent self, with beliefs about the world which make it simple, predictable and understandable for me. A belief that covid is a hoax and involved deception from many institutions, speaks to a higher-order belief that the world runs with deception everywhere (monological belief structure), which in turn speaks to a higher-order belief that the agent is smart and cognisant enough to understand why the world works the way it does. In tumultuous times such as during the coronavirus pandemic, there were a lot of complex changes to make sense of.

This also captures how conspiracy theories can be said to express certain values, whether they be political values or more personal values. Cassam (2019), for example, describes in detail that conspiracy theories are essentially forms of *political propaganda*, often espousing anti-semitism. Understanding conspiracist claims as confabulations, and therefore as claims which are motivated by a desire to express and signal one's values and treasured self-concepts, explains why and how one's political affiliations (especially if they are strong, which Van Prooijen et al (2015) found to predict belief in conspiracy theories) can bring the agent to adopt conspiracy theories which also embody those political values. It is a way for those values to be attributed to themselves and expressed. There is no *explicit* motivation to do these things – agents instead take themselves to be getting something right about the world, rather than choosing explanations about the world on the basis of their treasured self-concepts and values. This is just how confabulation works.

Furthermore, this illuminates how conspiracist claims could be confabulations even though confabulation is commonly seen as a more short-term, situation-specific phenomenon. The participants in Nisbett and Wilson's (1977) study, for example, give a reason for having chosen their stocking, and then presumably move on with their lives. Adherence to conspiracy theories, on the other hand, seems more like an ongoing project which structures much more of the agent's life, and mental life[5]. However, although specific instances of confabulation may be quite clearly circumscribed, the themes which are drawn on over time are constant. All the individual

[5] Confabulation might also be commonly thought of as a very individual phenomenon, but confabulations draw on 'good' reasons which successfully justify some attitude or behaviour *by the standards of the culture the agent is in*. In other words, the specific content of what justifies something, is drawn from the shared, surrounding environment.

instances of confabulation are still painting one picture; of the agent being coherent and understanding the world, and as having the chosen values of the agent (such as being kind, right-wing, a fan of horror movies, whatever). It is through this mechanism that confabulation can play a long-term role in determining future behaviour, to be in line with our imagined and idealised selves (De Bruin & Strijbos 2020).

In this way, many conspiracist claims can be understood within the framework of Mendalbaum's (2019) 'psychological immune system'. This kicks into action when we are threatened (even if we may not realise it consciously). He says "Among whatever other laws there are about belief change, we have reason to believe that there is a basic psychological immune system at work, constantly adjusting beliefs to ward off serious threats to one's sense of self." (2019, 152). And recently, Van Prooijin (2022) has also suggested that these kinds of processes could be in play in adhering to conspiracy theories. He says "conspiracy theories have psychological benefits by imbuing perceiver's worldview with meaning and purpose in a rewarding manner. Conspiracy theories enable an alternative reality in which perceivers (a) can defend a fragile ego by perceiving themselves and their groups as important, (b) can rationalize any of their beliefs and actions as legitimate" (2022, 1). Understanding conspiracist claims as confabulations helps us see quite how far these cognitive mechanisms can go in bringing agents to believe such bizarre and unreasonable things.

2.3. Feature three: Endorsing contradictions.

The third and final feature I will discuss is that this tendency to draw on further conspiracies even extends so far as endorsing contradictory conspiracy theories. Conspiracy theorists tend not only to subscribe to multiple conspiracy theories, *but even contradictory ones.* Wood et al (2012) found that the more that participants believed that Princess Diana had faked her own death, the more they believed that she had been murdered. And, the more they believed that Osama bin Laden was already dead when the US Special Forces found him, the more they believed that he is actually still alive. In our example of Robin, she claims that the coronavirus is just the flu, exaggerated by politicians, but also that it is very much a real and harmful weapon released maliciously by China. Adherence to contradictory conspiracy theories about coronavirus has been found; Miller (2020) found that a host of contradictory conspiracy theories about coronavirus were all predictors of each other. Three contradictory theories which were significant positive predictors of each other were: the virus was accidentally released by China; the virus was accidentally released by the U.S.; and the virus is a biological weapon intentionally released by China (2020, 5). This is one of the most worrying, bizarre and irrational features of

claims by conspiracy theorists, and is a something which prompts us to think that pathology must be involved on the part of the agent. However, as I will explain in section 3, this is not necessarily the case.

However, understanding these claims as confabulations helps again. In elaborating on our confabulations in response to pressure from others, we can easily run into contradictions. And we see this because we are, as discussed in section 1, we are primarily (though not solely) trying to provide motivating reasons which *justify*, rather than merely retrace a causal explanatory story of events. It's much easier to see how contradictory claims can end up being endorsed in the context of a search for justification, rather than a search for one, true causal history.

This floundering for further justifications when initial explanations are challenged or rejected is demonstrated clearly in Haidt's (2001) well-known incest study. They provided participants with a vignette about a brother and a sister who decide to sleep together. They use contraception and their relationship isn't negatively affected afterwards. Participants are asked whether what the siblings did was wrong and why they think that. Often, they would say it was wrong and then give reasons which referred to factors which the vignette actually explicitly ruled out as not applying. For example, they would say that the siblings risked an abnormal pregnancy, but the researcher would remind them that the vignette notes that they use contraception. Then, they might say that they could have ruined their good relationship. But again, the researchers remind them that the relationship was unaffected. As Bergamaschi Ganapini (2020) points out, the fact that participants 'chop and change' their story in response to the investigators suggests that they are seeking a justification of their position rather than a causal explanation of how they came to it. If it were the latter, we would surely see participants, in their responses, recognising that they may have made a mistake about the facts of the matter, but that it's nevertheless still how they came to their conclusion (2020, 195). Instead, participants are eager to reach for a new justification. But the participants can't have concluded that the act was wrong because of the risks of pregnancy, *and* not because of pregnancy risks because contraception was used, *and* actually because of the risks to the relationship, *and* not, and so on. Taken as explanations which seek to capture the causal history of their decision, these stories can't all be true. But taken as justifications, these statements do not contradict each other in the common aim of demonstrating that the agent is rational, trustworthy, cognisant of the reasons for their own behaviour. It is easier to see how contradictory statements (in the explanatory sense) can come out here, when the aim is to 'save face' socially and try out multiple justifications which explain a more basic sense or intuition. Rather than, trying to capture the one, true causal history of how the choice/attitude came to be. This focus on justifying and 'saving face' in the social world, instead, is the 'argumentative reasoning mechanism' described by Mercier & Sperber (2011). On

their view, producing and evaluating arguments has the function of learning about who to trust and how to gain others' trust, which is different to simply trying to close in on the truth of the matter.

Another, more familiar occurrence of this kind of contradictory thinking can be found in agents endorsing contradictory claims in the context of stereotypes about others (Wood et al 2012, 768). Marginalised groups, such as Jewish people and immigrants, are often chastised both for 'taking people's jobs' and 'stealing culture', whilst at the same time being 'layabouts receiving state benefits' and 'not bothering to integrate'. As causal explanations, these claims are clearly contradictory. But in terms of overarching values and a justification of racist sentiment, they are perfectly consistent. And, they could perfectly well all be given, by the same agent, as confabulations if that agent is challenged on why they support some racist policy, or why they displayed prejudiced behaviour. We see that contradictory and prejudiced attitudes like these are persistent.

Note that the participants in Nisbett and Wilson's (1977) study are unsatisfied with the explanation that certain stockings are chosen due to position effects, because it doesn't *justify* their choice well, and so they reject it. But Haidt's (2001) participants do also at least respond to challenges that their explanations are *explanatorily* inadequate (because they get the facts wrong), and search for other explanations for their stance. Given that agents respond both to challenges that their explanations are inadequate at justifying or explaining, this shows that they care about both things and that both things are in play in confabulation. By seeing these claims as confabulatory, this explains why conspiracy theorists (keenly) elaborate on their theories when they are pressed for being explanatorily inadequate, but are also strongly motivated enough to justify themselves that contradictory conspiracy theories get endorsed because they all contribute to that central interest, of emphasising their sense of self as coherent and understanding the world.

2.4. On 'filling a gap'.

I am not proposing that conspiracist claims and confabulation merely have the same structure or psychological architecture behind them, but further that conspiracy claims *are* confabulations. That is to say, that (often) conspiracy theorists are confabulating when they make conspiracist claims. So, more specifically, these conspiracist claims are a subset of confabulations. I have said that confabulations fill 'cognitive gaps'. There is a 'cognitive gap' in conspiracy theorists which we see is filled with conspiracist claims and confabulations. A number of what we could call causal psychological reasons for embracing conspiracy theories

have been identified and are discussed in a review by Douglas et al (2019, 7-9), and I list a number of them below:

- Consistently seeking patterns and meaning in the environment (van Prooijen, Douglas, & de Inocencio, 2018);
- Overestimating ability to understand complex phenomena (Vitriol & Marsh, 2018);
- A high need for cognitive closure (Marchlewska, Cichocka, & Kossowska, 2018);
- Lower levels of analytic thinking (Swarmi et al, 2014);
- Hypersensitive agency detection (Douglas et al, 2016);
- Unfulfilled existential needs (e.g. for agency and control) (Douglas et al, 2017);
- Feelings of low sociopolitical power and alienation (Bruder et al, 2013);
- A need to feel unique (Imhoff & Lamberty, 2017);
- Being defensive about one's social group, especially when that group is seen as 'low-status' and attacked (Uscinski & Parent, 2014);
- Having a 'conspiracy mindset', which, broadly, is a strong political and ideological mindset which distrusts powerful groups in society (Imhoff & Bruder, 2014).

It's very hard to imagine Robin drawing on any of these factors in her explanation of why she has the conspiracist beliefs which she has. In the same way, we didn't see any of the participants in Nisbett and Wilson's (1977) study identify the positioning of the stockings as a reason for their choice, and we didn't see any of the participants in Haidt's (2001) incest study say that a basic and unexamined, but deeply evolved, gut feeling of disgust is the reason for their moral judgement. There is a cognitive gap here which is filled with confabulation, in these participants' cases and in Robin's case. They use motivating reasons to justify their choice but they nevertheless do want an *explanation* of events as well. Hence, Robin draws on conspiracies which seem plausible to her and the participants draw on what they see to be rational and smart reasons to choose as they did.

Van Prooijin and Böhm (2023) have suggested that *feelings* of vaccine hesitancy might *precede* accepting anti-vaxx conspiracy theories, and conducted a study which provided some evidence for this from the temporal effects found. In essence, they found that people may be rationalizing their hesitancy around vaccines later on, with conspiracy theories regarding vaccines. So, these conspiracy theorists could be in a similar position to Haidt's (2001) participants, who have a relatively unexamined but strong felt feeling about something (disgust at incest; fear of vaccination), and come up with reasons for those sentiments after the fact which justify them in ways which preserve their senses of being rational and smart. Similar feelings might have been uncertainty during the coronavirus pandemic, or alienation within one's sociopolitical world. Once we are 'provoked' to examine these feelings, we'll be motivated to go for explanations

which fulfil other epistemic needs for us, like protecting our self-concepts and sense of agency, and this is what we see in cases of confabulation.

The fact that confabulators are filling a cognitive gap is something which means they are not aiming to deceive others. They are not aware of the shortcomings of their explanation, and that there are other, more accurate explanations which capture factors which were a strong influence on their behaviour. This brings a boundary to my claim about which conspiracist claims are confabulations, as it is in fact not *all of them*. Crucially, once there is any intention to deceive, there is no confabulation anymore. If someone knows that the conspiracy theory(ies) that they are advocating isn't true, or is very poorly supported, but they continue to promote it for other reasons (such as political gain, for example), then they are not confabulating even though they may be thought of as conspiracy theorists. A distinction which this is likely to track is that of the producers and the consumers of conspiracy theories (Cassam 2019, 33). It's widely accepted that different motivations are likely to be in play across these groups. The producers of conspiracy theories are more likely to be acting from malevolent intentions and sowing misinformation on purpose, whereas consumers are less likely to be doing this and therefore more likely to be confabulating. This is not a hard and clear divide though, and there are no doubt some actors who are difficult to place. For example, 'trolls' who are most likely 'bullshitting' with no regard for the truth either way.

3. The source of pathology.

Throughout the paper, I have been talking about everyday confabulation, as studied in normal, healthy subjects. I hope to have shown how even the most bizarre features of conspiracist claims, such as their bizarreness, stark ill-groundedness, prioritisation of self-concept over evidence, and violation of rationality in endorsing contradictions, can all in fact be explained with the everyday psychological apparatus behind confabulation and within us all. In most instances, then, I do not think we need to posit pathology in agents who subscribe to conspiracy theories.

However, confabulation is not only seen in the healthy population. It is also seen in patients as part of psychiatric illnesses (this is where confabulation was first studied). Patients with conditions such as dementia, Kosakoff's syndrome, anosognosia often present with clinical confabulation (Hirstein 2005, 8). In these cases, they fill cognitive gaps which arise from their psychiatric illness. In healthy people, the cognitive gaps are due to normal cognitive processing. In particular, any processing associated with 'type 1' processing (Kahneman 2011). For example,

gut feelings, intuitions, associations, motivated reasoning to avoid uncomfortable truth, and 'eureka' moments, all with little introspective accessibility.

Whether a case of confabulating is clinical or non-clinical, depends on the nature of the cognitive gap being filled. In cases of Alzheimer's, for example, confabulations are filling cognitive gaps in memory caused by the disorder. Hence, clinical confabulation. In everyday settings, confabulations fill gaps more commonly found in our cognitive architecture; system 1 thinking, motivated reasoning, biases etc. This means that whether a conspiracist claim is part of pathology or not, does not depend on any of the things which often prompt us to suspect that pathology must be present; bizarreness, ill-groundedness, irrationality. We already see these things in everyday confabulation, as I have shown. The nature of the pathology, if present, rather is parasitic on the type of confabulation we have. The pathology is 'upstream' of the anything about the contents of the conspiracist claims, despite how bizarre and ill-grounded they are. After all, some instances of clinical confabulation, where pathology is present, can actually be *less* bizarre. When someone with anosognosia is prompted to explain why they aren't moving their right arm (which they are unable to do, due to paralysis), they might say "I just don't feel like it, doctor" (Hirstein 2005).

This means that there isn't a scale of 'non-pathological' to 'pathological' along which, as claims get more bizarre and poorly supported by evidence, we are more likely to have pathology. Instead, we should look to the formation processes and nature of the cognitive gap being filled. Those same contents of some conspiracist claims, for example, that the government are covering up the presence of aliens on Earth, could be found within a framework of everyday confabulation or of clinical confabulation in a patient with a pathological condition, such as an elaborated delusion. This makes the matter of conspiracist claims manifesting pathology or not, a more trivial matter, which in fact just hinges on the type of confabulation, which depends on the aetiology and belief formation processes of how the agent came to espouse such views[6].

4. Conclusion

To conclude, I have outlined that when people confabulate, they make ill-grounded statements which express motivating reasons which aim to justify their positions. This can explain the belief formation of conspiracy theorists. Conspiracist claims are defensive and self-sealing with counterevidence being 'waved away', which we see in secondary confabulation. This is often done with reference to further conspiracies, and I have argued that understanding these claims

[6] See Bortolotti 2023 for a closer look at the cognitive processes behind conspiracist ideation, for possible pathology. Similarly, she argues that these processes are not pathological.

as confabulations highlights the monological structure of this; all these claims are made in aid of a higher-order, overarching belief that the agent has a smart understanding of the world and their place in it. Positing deception everywhere is one way to do this. Finally, conspiracist claims are often contradictory. But, taken as confabulations which aim to *justify*, they do not really contradict each other in terms of trying to justify the overarching higher-order belief that, for example, organisations cannot be trusted, and that the agent has a good understanding of the world.

Therefore, the features of conspiracist claims which often prompt us to think that conspiracist claims may signal pathology on the part of the agent, ought not. These are features such as being bizarre, ill-grounded, lacking in evidence (due to prioritising self-concepts), and being unreasonable to the point of being contradictory. This doesn't make pathology within the agent *impossible*, but these features of the contents of conspiracist claims are not the place to look. Instead, one should look at the aetiology and nature of the cognitive gap being filled.

Bibliography:

Bergamaschi Ganapini, M., (2020), 'Confabulating reasons', in Topoi, 39(1), pp.189-201.

Bortolotti, L., (2018), "Stranger than fiction: costs and benefits of everyday confabulation", in Review of philosophy and psychology, 9(2), pp.227-249.

Bortolotti, L., (2020), *The epistemic innocence of irrational beliefs*, Oxford University Press.

Bortolotti, L. and Cox, R.E. (2009), 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. Consciousness and Cognition, 18(4), pp.952-965.

Bortolotti, L. and Murphy-Hollies, K., (2022), 'Exceptionalism at the time of Covid-19: Where nationalism meets irrationality', in *Danish Yearbook of Philosophy*, 55(2), pp.90-111.

Bortolotti, L., (2023), 'Is it pathological to believe conspiracy theories?', in *Transcultural Psychiatry*, 0(0).

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013), 'Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire', in *Frontiers in Psychology*, 4(225).

Cassam, Q., (2019), *Conspiracy theories*, John Wiley & Sons.

Coltheart, M., (2017), 'Confabulation and conversation', in Cortex, 87, pp.62-68.

de Bruin, L. and Strijbos, D., (2020), 'Does confabulation pose a threat to first-person authority? Mindshaping, self-regulation and the importance of self-know-how', in *Topoi*, 39, pp.151-161.

Douglas, K. M., Sutton, R. M., Callan, M. J., Dawtry, R. J., & Harvey, A. J. (2016), 'Someone is pulling the strings: Hypersensitive agency detection and belief in conspiracy theories', in *Thinking & Reasoning*, 22(1), 57–77.

Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017), 'The psychology of conspiracy theories', in *Current Directions in Psychological Science*, 26(6), 538–542.

Douglas, K.M., Uscinski, J.E., Sutton, R.M., Cichocka, A., Nefes, T., Ang, C.S. and Deravi, F., (2019), 'Understanding conspiracy theories', in *Political Psychology*, 40, pp.3-35.

Evans, G., 1982, *The Varieties of Reference*, Oxford: Oxford University Press (ed. J. McDowell).

Haidt, J. (2001) 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement', Psychological Review, 108(4), pp. 814–834.

Hall, L., Johansson, P., Tärning, B., Sikström, S. and Deutgen, T., (2010), 'Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea', in *Cognition*, 117(1), pp.54-61.

Hirstein, W. (2005). Brain fiction: Self-deception and the riddle of confabulation. MIT Press

Ichino, A. and Räikkä, J., (2020), 'Non-doxastic conspiracy theories', in *Argumenta*, pp.247-263.

Imhoff, R. and Bruder, M., (2014), 'Speaking (un–) truth to power: Conspiracy mentality as a generalised political attitude', *European Journal of Personality*, 28(1), pp.25-43.

Imhoff, R., & Lamberty, P. K. (2017), 'Too special to be duped: Need for uniqueness motivates conspiracy beliefs', in *European Journal of Social Psychology*, 47(6), 724–734.

Jongepier, F., & Cassam, Q., (2020), 'Radically self-deceived? Not so fast', in Peels, R., de Ridder, J., & van Woudenberg, R., (eds.), Scientific Challenges to Common Sense Philosophy, Routledge.

Johansson P, Hall L, Sikström S, Olsson A (2005), 'Failure to detect mismatches between intention and outcome in a simple decision task' in *Science*, 310:116–119

Johansson P, Hall L, Sikström S, Tärning B, Lind A (2006), 'How some-thing can be said about telling more than we can know', in *Conscious Cognit* 15, pp.673–692

Keeling, S., (2018), 'Confabulation and rational obligations for self-knowledge', in *Philosophical Psychology*, 31(8), pp.1215-1238.

Klein, C., Clutton, P. and Polito, V., (2018), 'Topic modeling reveals distinct interests within an online conspiracy forum', in Frontiers in psychology, 9, p.189. Available here: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00189/full <Accessed 05/2022>

Lewandowsky, S. and Cook, J., (2020), *The conspiracy theory handbook*, available at http://sks.to/conspiracy [accessed August 2023]

Malle, B.F., Knobe, J.M. and Nelson, S.E., (2007), 'Actor-observer asymmetries in explanations of behavior: New answers to an old question', in *Journal of personality and social psychology*, 93(4), pp.491-514

Mandelbaum, E., (2019), 'Troubles with Bayesianism: An introduction to the psychological immune system', in *Mind & Language*, 34(2), pp.141-157.

Marchlewska, M., Cichocka, A., & Kossowska, M., (2018), 'Addicted to answers: Need for cognitive closure and the endorsement of conspiracy beliefs', in *European Journal of Social Psychology*, 48, 109–117.

Mercier, H. & Sperber, D., (2011), Why do humans reason? Arguments for an argumentative theory', in Behavioural and Brain Sciences, 34, pp.57–74

Miller, J.M., (2020), 'Do COVID-19 conspiracy theory beliefs form a monological belief system?', in *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2), pp.319-326.

Moscovitch M (1995) Confabulation. In: Schacter D (ed) Memory dis- tortion. Harvard University Press, Cambridge, pp.226–251

Murphy-Hollies, K., (2022), 'Political Confabulation and Self-Regulation', in *Royal Institute of Philosophy Supplement*, 92, pp. 111-128

Nisbett, R. E. and Wilson, T. D. (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes', in Psychological Review, 84(3), pp. 231–259

Örulv, L. and Hydén, L.C., (2006), 'Confabulation: Sense-making, self-making and world-making in dementia', in Discourse Studies, 8(5), pp.647-673

Sandis, C., (2015), 'Verbal reports and 'real' reasons: Confabulation and conflation', in *Ethical Theory and Moral Practice*, 18(2), pp.267-280.

Stammers, S., (2020), 'Confabulation, explanation, and the pursuit of resonant meaning', in Topoi, 39(1), pp.177-187

Strandberg, T., Sivén, D., Hall, L., Johansson, P. and Pärnamets, P., (2018), False beliefs and confabulation can lead to lasting changes in political attitudes, in *Journal of Experimental Psychology: General*, 147(9), pp.1382-1399

Sullivan-Bissett, E., (2015), 'Implicit bias, confabulation, and epistemic innocence', in *Consciousness and Cognition*, 33, pp.548-560.

Swami, V., Chamorro-Premuzic, T., & Furnham, A. (2010), 'Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs', in *Applied Cognitive Psychology*, 24, pp.749–761.

Swami, V., & Coles, R. (2010). The truth is out there: Belief in conspiracy theories. The Psychologist, 23, pp.560–563

Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S. and Voracek, M., (2011), 'Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories', in *British Journal of Psychology*, 102(3), pp.443-463.

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014), 'Analytic thinking reduces belief in conspiracy theories', in *Cognition*, 133(3), 572–585.

Uscinski, J. E., & Parent, J. M. (2014), *American conspiracy theories*, New York, NY: Oxford University Press.

Van Prooijen, J.W., Krouwel, A.P. and Pollet, T.V., (2015), 'Political extremism predicts belief in conspiracy theories', in *Social psychological and personality science*, 6(5), pp.570-578.

Van Prooijen, J.-W., Douglas, K., & De Inocencio, C. (2018), 'Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural', in *European Journal of Social Psychology*, 48, 320–335

Van Prooijen, J.W., (2022), 'Psychological benefits of believing conspiracy theories', in *Current Opinion in Psychology*, 47.

van Prooijen, J.W. and Böhm, N., (2023), 'Do Conspiracy Theories Shape or Rationalize Vaccination Hesitancy Over Time?', in *Social Psychological and Personality Science*, 0(0).

Vitriol, J. A., & Marsh, J. K. (2018), 'The illusion of explanatory depth and endorsement of conspiracy beliefs', in *European Journal of Social Psychology*, 48, 955–969.

Wood, M.J., Douglas, K.M. and Sutton, R.M., (2012), 'Dead and alive: Beliefs in contradictory conspiracy theories', in *Social psychological and personality science*, 3(6), pp.767–773.

Confabulation and Reasons for Love.

Kathleen Murphy-Hollies (University of Birmingham) [Corresponding author]
Pilar Lopez-Cantero (Tilburg University)

Introduction

Most of us believe that love for our partners and friends is not random: we love *them* and not other people, and our love for them has to do with something about them specifically. Although people do not often ask themselves why they love their partners and friends, on occasion an answer to this question is prompted. For example, friends may be unable to understand one's partner choice - "what do you see in him?". Other times, it is the loved person who asks - out of curiosity or need for self-affirmation. It is also common to wonder about the appropriateness of one's love in times of relationship crisis, or in the aftermath of a break-up - "I still love her after all the harm she's done, but why?". These are all ways of asking about the reasons for love, so we consider that the individuals in these examples are occupying a *reason-giving stance* with respect to their love. These circumstances are often highly pressured; we're keen to be able to give an answer for why we love the people we love, to others or to ourselves. Philosophers of love are sceptical as to whether we can give normative reasons for love when we find ourselves in the reason-giving stance.

In this paper, we refuse this pessimist view, which we believe relies on a conception of people as perfect epistemic agents. We argue that people do and can give reasons for their love in order to *feel* justified, and that these reasons can eventually become, in fact, normative reasons for love. This epistemic practice does not follow the steps that a perfect epistemic agent would take (to look at the evidence, assess it and provide with reasons); in fact, these reasons are obtained through confabulation. We suggest that agents are likely to confabulate reasons in these reason-giving stances because this is what agents tend to do when they are prompted to give reasons for their behaviour, choices or attitudes etc, but do not have a comprehensive understanding of all the factors in play. So, although real agents are epistemically limited and unlikely to be able to give a comprehensive answer, they are unlikely to say nothing at all. As we will explain, however, these confabulated reasons are epistemically problematic in a few ways - they are post-hoc, ill-grounded, and often consist of false statements. We also highlight how paying closer attention to what real agents actually do and say when in these reason-giving stances complicates the long-standing split found in the literature between rationalist and anti-rationalist accounts of love.

The argument goes as follows. In Section 1, we introduce the debate on the rational justification of love as appropriateness, and explain the 'gaps' we have with respect to our epistemic access to the normative reasons for our love. In Section 2 we introduce the concept of confabulation, and show that people in the reason-giving stance often confabulate. Confabulating people might: (i) give reasons that are false, insofar they do not correspond with facts of the world. Or, if they are not false, they are still (ii) ill-grounded, in that even if they correspond with facts of the world, they are epistemically faulty in some other ways. In these contexts, this means that those reasons (even if true) are not capturing the normative reasons for that love (so in other words, the individual is not loving on the basis of that reason). In Section 3 we argue that when confabulating, people are aiming at *feeling* that their love is justified. This sense of justification as felt is not completely divorced from the sense of justification as appropriateness: on the contrary, aiming at felt justification can result in the reasons provided becoming appropriate in the traditional sense. We describe that two possible ways that this can happen in section 4: Articulating confabulated reasons results in false claims becoming true (i.e., articulating the reasons changes the facts in the world), or such explicit consideration and endorsement of a feature of the world picked out as a confabulated reason means that that reason becomes normative. In section 5 we explain how our view aligns with a recent trend in the philosophy of love to obscure the divide between Rationalism and Anti-Rationalism, with the advantage of overcoming the assumption of a perfect epistemic agent. Finally, in section 6, we address the worry that confabulation could enable harmful and abusive relationships.

1. Reasons for Love and Epistemic Access

Philosophers have long concerned themselves with the question of whether there are facts of the world that justify our love for particular people. Traditionally, they have been divided in two camps, Rationalists and Anti-Rationalists. Rationalists believe that love is the sort of phenomenon that can be justified in terms of reasons, while Anti-Rationalists reject this claim.

Rationalists consider love an appraisal of value, which then can be justified in terms of reasons. They disagree among themselves on which kind of reasons these are—for example, intrinsic properties of the loved person, or 'personal qualities' (Keller 2000), relational properties like shared history (Kolodny 2003), or a mix of both (Jeske 2007, Protasi 2016, Naar 2017). But they all agree on the sentiment that love "happens for better or worse reasons" (Helm 2010: 22). At the core of their view, then, is the matter of which reasons are in fact appropriate or fitting reasons for love. Within the

debate, the terms 'appropriateness' and 'fittingness' tend to be used interchangeably—for reasons of clarity we will use the former only from now on. In turn, the notion of appropriateness is used in three different ways: appropriateness can refer to whether love is prudentially or morally desirable; to whether reasons for love are apt; or to whether love is rooted on evidence. In the first sense, love can be appropriate if it leads to flourishing, and inappropriate if it is damaging to oneself or others.

In the second sense, love can be inappropriate if it is not rooted in the personal qualities of the beloved: for example, "I love my friend because she has a lot of money" or "I love my husband because of the shape of his feet". Not any personal quality is an apt reason for love, given the shared assumption by philosophers of love that people should be loved *for their own sake*. For Rationalists, to love someone for their own sake means to *value* them *as the particular people they are*, or in other words, caring for *their identity*. For that reason, only qualities that are a fundamental part of the beloved's identity count as potential reasons for love (Helm 2010, Delaney 1996). Anti-Rationalists also share the view that people should be loved for their own sake. For them, this means to be *concerned* for the beloved's identity as for one's own identity (Frankfurt 2004: 83).

In the third sense, love is appropriate if it succeeds in picking up actual features of the world. For example, many cases of unrequited love where there is imagined reciprocity, overvaluation or transference are inappropriate in this sense ("She said hello because she really cares about my well-being" or "I have not met a gentler soul in my entire life" said about someone who beats animals). These are different ways that Rationalists have tried to distinguish between better and worse reasons for loving a particular individual.[1] Here, we treat each of these three senses of appropriateness as sufficient: if we love for good reasons in any of the three senses, love is appropriate, and thus love is *justified*.[2]

Throughout this paper, we will refer to reasons that justify love as 'normative reasons'. Again, it is important to insist on the fact that the property of being *good* reasons is not taken to mean 'good' or

[1] For an explanation of the first and the second sense or appropriateness, see Han (2021: 4-6). The third sense is implicitly assumed in the discussion, as evidenced by the widespread worry that love may be grounded on imagined qualities–see (Jollimore 2011: 7) for a discussion and a refutation of this worry.

[2] This stance would be rejected by those that argue that the only personal qualities that can justify love are moral qualities (e.g. Velleman 1998, Abramson and Leite 2011, Setya 2014), and by those who argue that love for immoral people is inappropriate (e.g. Elder 2014, Isserow 2018, Mason 2022). We accept that incompatibility, since in our view it's been sufficiently argued that both these claims are false–see e.g. Pismenny (2021) for a rejection of the former and Cocking and Kennett (2000) for a rejection of the latter claim.

'normative' exclusively or mainly in the moral sense. That is only one of the ways of understanding appropriateness–the first sense we just described. For example, if a chronically online misogynist loves Jordan Peterson because Peterson supports patriarchy, that is a bad reason in the first, moral, sense. However, this love is still appropriate in the second sense: it is apt, since one of Peterson's personal qualities is that he supports patriarchy. Supporting patriarchy is part of who he is. Love for Peterson would be inappropriate in the second sense if it were based on some contingent property of Peterson, like his hair colour. And it would be inappropriate in the third sense if the misogynist loved Jordan Peterson for his sobriety, given that Peterson is simply not sober and his drug addiction is well-known, so love would not be picking up actual facts of the world. Normative reasons for love refer to facts of the matter with regards to a specific instance of love, which render that love justified, or appropriate.

On the opposite side of the debate, Anti-Rationalists deny that love is the kind of phenomenon that can be justified in terms of reasons. Harry Frankfurt (2004) is a renowned proponent of this approach (see also Thomas 1991, Zangwill 2013). For Frankfurt, love is an expression of care, not a response to reasons. In turn, care is a mode of disinterested concern for the loved person's sake, and hence not justifiable in terms of her intrinsic or relational properties. We do value qualities such as their character or our shared history, but this is a bestowal, and not an appraisal, of value.

On this basis, it seems obvious enough that Rationalists may claim that when placed in a reason-giving stance, people will give normative reasons for love. According to the Rationalist view, if we were to ask a friend why they love their spouse, they would give answers of the sort: "Because she is funny and charming", or "Because we have been through so much together". Although Anti-Rationalists believe that love is not justified in terms of reasons, that does not entail that people cannot give a response when prompted. They can accept reasons in the shape of coarse explanations, such as "She is my best friend" or "I just love him". When considering the reason-giving stance, then, the main difference between Rationalists and Anti-Rationalists is that the former believe that we can give normative reasons, while the latter believe that we can only give explanatory reasons (which tend to be wide and not particularly informative about reasons for loving that particular individual).

Something that both sides of the debate have in common, however, is that they both completely overlook the matter of epistemic access. This is the question of how people can know these reasons, if at all. The thought is not whether there are normative reasons for love or not, but whether we can plausibly claim to have epistemic access to *any* reasons for love. Philosophers of love have mostly been silent on this matter, which reveals an implicit association of the reason-giving stance with a

perfect epistemic agent, who is able to access and articulate any kind of reasons that in fact exist (normative or explanatory). The only discussion of epistemic access to the reasons for love has recently been put forward by Hichem Naar (2022: 144), who has a rather pessimistic outlook: reasons for love are "difficult to articulate, if not completely inarticulable". He further develops this claim:

> If asked why you love someone, …it will be difficult, if not impossible, to say exactly what your reason is… At best, one can employ rough characterization such as 'She is just her', 'He is the subject he is', 'They are a world', but these won't surely give others a precise idea of your reason. Testimony, therefore, is severely limited when it comes to our epistemic access to reasons for loving particular individuals; in fact, it seems that testimony by itself won't allow us to come to possess a reason to love an individual. (Naar 2022: 144).

In other words, Naar denies that when we ask our friend why they love their spouse, our friend will be able to offer normative reasons for their love. They will only be able to give generic explanatory reasons. It follows then that the only possible result when we occupy the reason-giving stance is that we cite imprecise reasons that are not the normative reasons for our love. This leads Naar—who is a Rationalist—to argue that normative reasons for love can only be acquired through first-hand experience of the object of love, and to endorse a perceptual account of love's rationality. We will not explore the potential merits and limitations of Naar's perceptual view, since what interests us is to highlight how he assumes an 'all-or-nothing' epistemic approach with respect to normative reasons: since we do not have epistemic access to normative reasons for love, then we cannot articulate them. We can at most articulate coarse explanatory reasons, of the sort that Anti-Rationalist propose. Naar is by default endorsing that we are perfect epistemic agents, since the fact that we cannot access normative reasons means that we will not, in fact, give any normative reasons while in the reason-giving stance.

Although we agree with Naar on the epistemic obscurity of reasons for love, we disagree with him that 'testimony by itself won't allow us to come to possess a reason to love an individual', and we expand on this in section 4. We also reject the notion of the perfect epistemic agent and the subsequent all-or-nothing approach that is implicit in his account, and more widely in the Rationalist view that assumes full epistemic access to normative reasons for one's love. In reality, the reason-giving stance with respect to love is bound to happen in a less straightforward manner. We illustrate this with an example. At the end of *When Harry Met Sally*, Harry rushes to the New Year's party to tell Sally that he is in love with her. To his surprise, Sally is unimpressed with his big confession, and asks

him with scepticism, 'What is that supposed to mean?' as she begins to leave. Effectively, Sally puts Harry in the reason-giving stance, which Harry then follows with a more elaborate declaration:

> "How about this way. I love how you get cold when it's 62 degrees out. I love the way your mouth turns down just a little bit, right there. I love how it takes you an hour and a half to order a sandwich. I even loved when you used my sweater as a Kleenex. I love that after I spend the day with you, I can still smell your perfume on my clothes. And I love that you are the last person I want to talk to before I go to sleep at night. It took me eleven years to figure this out. And I came here tonight because when you realize you want to spend the rest of your life with somebody, you want the rest of your life to start as soon as possible".

Harry is openly admitting that he is far from a perfect epistemic agent: by his own account, it took him years to figure this out. Moreover, as external observers we may judge that some of these are probably not actual normative reasons for Harry's love of Sally. This can be cashed out through a counterfactual: if it would still be the case that an individual would love a particular person in absence of the set of properties S, then S does not constitute a set of normative reasons for love for that person. Maybe the individual offers S as a response to the questions above, but the reasons that justify their love are other properties of the loved person. It seems plausible here that Harry would still love Sally if her basal body temperature was higher, or if she became faster at ordering at bodegas. Nevertheless, when put in the reason-giving stance, Harry gave all these reasons—and Sally accepted them as apt reasons, since her scepticism was dismantled.

This is the phenomenon we are interested in: the fact that despite not being perfect epistemic agents, we do engage in reason-giving about our love when it becomes necessary. In other words, we change the question about the reasons of love from 'is love rationally justified?' to 'what do we do when we find ourselves in the reason-giving stance with respect to love?'. We start from the assumption that we are not perfect epistemic agents that have direct and clear access to the reasons for our attitudes and behaviours, but that, as with Harry, that does not stop us from giving such reasons. The mechanism that allows us to do so is *confabulation*, which we explain in the next section. Through the introduction of confabulation, we will show that in our testimony, we can provide others (and ourselves) with normative reasons for love.

2. Introducing confabulated reasons

We have established that we are interested in what people do when they have poor epistemic access to reasons for something, and in these cases, for loving someone. It is found that when people are prompted to give reasons for a feeling, choice, behaviour or attitude of theirs, lacking epistemic access to the actual reasons for that behaviour often doesn't stop them from giving reasons. So, individuals certainly know that they have some choice or attitude - in these cases, that they love someone - but they do not have access to the source of these attitudes - why they have them (Bortolotti 2009, 628; Haidt 2001; Lawlor 2003). People will often *confabulate* reasons for their attitude, choice etc. At a first gloss, this means that individuals will fabricate reasons on the spot for their attitude, but those reasons are epistemically faulty in a couple of ways which we explain below. So, in situations like the one described above, we suggest that people will often confabulate reasons for love when someone presses them on "what do you see in him?", for example. It's very important to note that confabulation is done *with no intention to deceive*, so confabulators believe that they are giving a genuine account of the reasons for their attitude etc.[3]

Confabulated reasons are produced *post-hoc*, and in these cases, upon being *provoked* (Bortolotti & Cox 2009, 953). This means that it is only upon being asked (i.e., only when placed in the reason-giving stance), that individuals put together reasons for their behaviour. As mentioned above, when agents have poor introspective access to the reasons for their choices or attitudes, they have a *cognitive gap* there which they fill with confabulation. Why are individuals so keen to fill these gaps with post-hoc fabricated reasons? In general, individuals are keen to be able to respond to social prompting for reasons, so that they can avoid the discomfort of being dumbfounded (Haidt 2001), and instead signal that they are rational and understand themselves (Bergamaschi Ganapini 2020). We are motivated to continue participating in the social interaction, and more generally to remain integrated in the social world (Bortolotti 2018a). We are also motivated to preserve and emphasise our positive self-concepts in our confabulations (Sullivan-Bissett 2015, 552).

Our view here is that there is a direct application of the epistemic practice of confabulation to the reason-giving stance about reasons for love. When Naar says that agents do not have epistemic access to the normative reasons of our love, we can understand that as saying that agents have a cognitive gap, to which direct access is unsalvageable. However, Naar does not take into account that even if that were true, that does not prevent people from attempting to fill this cognitive gap, and subsequently providing reasons for love via testimony. Love is a highly charged topic, emotionally and socially. The pressures to 'save face' and give what are perceived to be 'good' reasons for loving the

[3] This is always the case, because confabulators are neither liars nor bullshitters.

person you love, are strong. Here, 'good reasons' should be understood in a fourth sense that has nothing to do with the three senses of appropriateness described above, but rather as something which provides the individual with good *social justification*. It is difficult and awkward to embrace not having good reasons for why you love someone; why, for instance, you feel so strongly about someone, spend so much time with them, think about the relationship so much, sacrifice a lot for that person, and so on. Given these considerations, and to protect positive self-images, people are unlikely to invoke loving someone for something like their money: it would look morally callous, and it is not a socially well regarded reason for loving a particular individual. All of this means that despite an agent's poor epistemic access to the reasons for their love, they formulate some in order to be able to respond to this social prompting.

Here is an example. Imagine that Zara is asked why she loves Chris. Their romantic relationship has been running smoothly for some time, but Zara hadn't much considered the reasons for her loving Chris until now, when she is prompted by a friend. She says 'I love how generous he is, and we have such a good time together in the evenings after work'.

Zara technically gives two reasons for loving Chris, and we use each to tease apart a distinction between false and ill-grounded confabulated reasons. Firstly, Zara says that one reason she loves Chris is that he is generous. Imagine that this is false; Chris is clearly not generous. He is stingy, and always keeps a keen protective eye on all of his possessions. However, Zara has always thought that generous men are good men, and being in a relationship with a generous man has just always been part of her conception of what a loving relationship consists of. So, this biases some of her memories of Chris. She also, secondly, says that the quality time they spend together on the sofa after work is a reason she loves him. Imagine that this reason is *ill-grounded*. This means that although the statement may be true about the world (they do indeed spend the pleasant evenings together on the sofa), this reason does not play the role which Zara thinks it does in justifying that love. A way of understanding this is that the counterfactual we gave in section 1 is not met; if they were to stop spending their evenings together on the sofa, then Zara would actually continue to love Chris. And so it is not playing the important role of justifying that love.

In the confabulation literature, confabulations are said to be ill-grounded in order to capture the fact that they are not necessarily false, confabulations can be true, but only by accident or chance. The process by which agents give these reasons is still unreliable, and usually not a good response to the evidence (a confabulator might confabulate the right time when asked, but this is not based on the clocks around her). As described above, confabulation is motivated by desires to 'save face' and

preserve a coherent sense of self, and here Zara mentions their quality time on the sofa together as a reason for her loving Chris, but this is stemming from the fact that she she feels that this makes a sensible story for why she loves Chris. She does not realise this, of course, and believes her account to be totally authentic. But she does not capture the reasons that are actually justifying that love - in other words, the normative reasons for her love. In summary, confabulations are always ill-grounded and always post-hoc, they are *usually* false but this is not necessary.

With regards to making false statements, there is plenty of evidence to show that people's descriptions of their loved ones and relationships are often distorted, and that the reasons given for their loving someone often poorly match the details of the case (as we see in Zara's case when she falsely describes Chris as generous). Sandra Murray and colleagues have run a number of studies showing this. Murray et al (1996a) found that participants idealised both their relationships and their partners, taking "considerable license in constructing impressions of their partners. [...] [representations] appeared to reflect their tendency to see their partners as they wished to see them, through the filters provided by their ideals and rosy self-images" (92). These findings are often discussed in terms of people being *unrealistically optimistic*, or having an optimism bias, when it comes to judging their relationships and their partners. They also reflect superiority bias (rating our partners and relationships as better than most) and the love-is-blind illusion (failing to see our parters' faults and judging them as better than average) (Bortolotti 2018b). Reporting on the quality of our relationships and the qualities of our partners on the basis of biases is ill-grounded; it's likely to result in false statements (as seen here) and isn't a careful enough response to evidence. These kinds of statements can therefore be seen as confabulations, because the participants do not have introspective access to the effect of optimism bias on their appraisals, and we are motivated to have nice sense-making stories which validate our perceptions and our relationships[4].

Despite claiming that there are strong pressures present for confabulating in these contexts and also providing empirical evidence of it, we are not claiming that *all* reasons for love are in fact confabulated. Sometimes, people will not have cognitive gaps and will successfully identify the normative reasons for their loving someone. Having said this, it is surely rare for someone's loving of another to be a straightforward matter which they fully understand and have complete epistemic access to. In situations where we have complex phenomena, cognitive gaps and motivations to give good reasons (in the sense of social justification) for something, we are likely to see confabulation.

[4] See Örulv and Hydén 2006 for a discussion of the sense-making function on confabulation, particularly in cases of dementia.

In summary, bringing confabulation into this discussion shows that agents do give more detailed responses than suggested by Naar about their reasons for love. They just do not behave like perfect epistemic agents who would only give a detailed answer in the reason-giving stance; a response according to the facts that they have actual evidence for. They give an answer with detailed reasons, and these are confabulated reasons.[5]

At this point, a philosopher of love, might understandably say, 'so what?'. So agents struggle to give accurate answers about the reasons for loving the people they love, and instead confabulate. Philosophers of love are trying to figure out what (if any) *normative* reasons justify instances of loving people, independent of people's limited and imperfect appraisals of their circumstances. We may have shown that having a cognitive gap in the way portrayed by Naar does not stop people from giving a response that goes beyond coarse explanations, but we have not shown that the reasons that people give are in fact normative reasons. In the next section, we lay some groundwork for arguing in section 4 that confabulating can indeed provide the normative reasons for love which rationalists *are* interested in.

3. From explanation to *felt* justification

Confabulations stem from imperfect agents having limited cognitive resources and access. Confabulations, therefore, look like explanations which plug those gaps - explanations of one's circumstances and of one's own attitudes etc. In this section, we describe how actually, we are doing so much more than simply trying to give an *explanation* of ourselves in cases of confabulation.

In section 1, we explained that the difference between Rationalists and Anti-Rationalists lies in whether love can be justified with reasons. Anti-Rationalists think love is not the kind of phenomenon that can be justified with reasons, although they acknowledge that it can be *explained* with reasons. The difference is expressed in the following example raised by Aaron Smuts—an Anti-Rationalist:

> Consider the abused housewife who refuses to leave a violent brute who doesn't seem to care about her in the least. 'I love him', she says…We can imagine the frustrated friends of the abused housewife asking her to, in effect, justify her love. They do not merely expect an explanation. If

[5] It could be said here that in that case, Naar would at least be right to claim that agents will never give non-confabulated reasons. We accept that as a logical consequence of our claim, but this is not a problem for our argument, that does not hinge on our ability to offer non-confabulated reasons when in the reason-giving stance.

> that is all that they were after, they would be content with any old explanation, such as: 'He reminds me of my father whose approval I could never earn.' That surely wouldn't satisfy her friends. They would take that as cause for her to leave. They do not merely want an explanation, they want something that looks like a justification (Smuts 2014: 523-524).

Smuts offers this example as a potential objection to Anti-Rationalism, but we focus on how it is illustrative of the stark difference that philosophers of love make between explanation and justification. However, this distinction between justification and explanation may be challenged, and some metaphysical accounts do not consider them to be completely different relations (e.g. Väyrynen 2019). Here, we will not fully endorse the equivalence between justification and explanation, since that requires a complete departure from the debate–it would entail that the whole discussion between Rationalists and Anti-Rationalists is misguided.[6] Instead, we show that the boundaries between justification and explanation are more malleable than is assumed in the debate.

The reason why philosophers of love have not concerned themselves with the reason-giving stance is that it seems that giving reasons and having reasons are two completely different questions, where only the latter is connected to justification. When in the reason-giving stance, it may be thought that everyday individuals are aiming (but, in the case of confabulation, failing) to trace and give a much more causal, explanatory story than anything the philosophers of love are after. An individual might give an accurate causal explanation for love, which a rationalist might agree with, but these same reasons wouldn't be acceptable as the normative reason for that love. For instance, it may of course be very true to say that taking jobs at the same company and literally bumping into each other in the corridor is a causal explanation for Zara and Chris falling in love, but nobody is thinking that that these kind of explanatory reasons for their falling in love are the ones which make it justified.

We start with a clarification about confabulation, and then explain why philosophers of love should in fact be interested in confabulation in the reason-giving stance. Firstly, confabulation is not simply an (unsuccessful) attempt to trace an accurate causal story of how an agent's attitude etc. came about. We have already explained that individuals want to tell a *good story*; the key motivation of confabulation is to justify oneself in the face of some pressure from the social world. So, agents are

[6] That is, of course, a possibility, and one that we do not find completely unconvincing. However, that would be a matter for a different discussion. If it were true that the whole debate is misguided, our argument still makes a contribution with regards to the reason-giving stance, by highlighting how testimony and felt justification can contribute to shaping love for particular individuals.

actually already in the business of trying to justify themselves rather than simply explain themselves. Some perfectly good causal explanations are rejected by agents because they do not provide them with adequate justification, in their eyes. For instance, Bergamaschi Ganapini (2019) explains that experiment participants who were told that their choice of product from a line-up in a consumer survey was brought about by a simple right-side bias, rather than because of any particular feature of the specific product they'd chosen, rejected those explanations specifically because they did not provide them with a good justification. Others in one's social environment are unlikely to think that choosing a particular consumer product because it was on the right hand side successfully signals one's rationality and, if the subject matter were more morally charged, one's trustworthiness in moral matters (ibid, 193). This is not because we think confabulators are more likely to lie or distort important moral matters, but more simply that humans prefer to make connections with and embark on joint projects with others who are reliable, trustworthy communicators of facts of the matter[7]. In essence, when agents give inadequate causal explanations (and this is identified by others), they care about this in a way which is deeper than simply realising that another explanation must in fact explain what's happened here. And this is somewhat reflected in the example of the abused wife given above, discussed by Smuts, where the explanations which the woman is giving are not providing anyone present with adequate *justification*.

However, it is not impossible for causal explanations to provide agents with this kind of sought justification. Agents are most likely looking for reasons for their attitude or behaviour which function *both* as good causal explanations and justifications, as is suggested by Stammers (2020) who argues that confabulators seek to give explanations which are also imbued with 'resonant meaning' for them. This means that both elements are important to confabulators.

Furthermore, we find that different causal explanations are received as better or worse justifications over time. For instance, relatively recently we have seen a shift in embracing reasons which refer to things like one's attachment style, one's psychiatric diagnosis, even one's star sign, and so on, in justifying behaviour. These are technically causal reasons, because they would not be employed as the reasons which the agent is taking herself to be acting on - as needs to be the case for something to be a motivating reason. For example, an agent would never phrase this as '*my reason for* my string of failed relationships is my insecure attachment style', or 'the reason *I am acting on* in having a string of failed relationships is having an insecure attachment style'. Rather, the insecure attachment style

[7] This view is in line with Mercier and Sperber's (2011) account of human reasoning; that our argumentative reasoning mechanisms specialise in *producing arguments* with the aim of convincing others and bringing them to trust us, rather than in simply tracking truth.

is a more external, causal reason *that* someone has a string of failed relationships. Nevertheless, this can afford an agent justification. In essence, we are using confabulation to show that the divide between explanation and justification is much more porous than we might think, given that both are clearly in play with confabulated reasons.

However, philosophers of love may still be concerned that this is a different sense of justification than the sort they are interested in. The justification involved in confabulations, described above, is the contingent and social sense of justification which we described in the previous section. Whereas, philosophers of love are after normative reasons which justify love by less contingent way. The basis of whether some shared reason 'works' as a justification for an individual depends on whether it's *working in one's current social world* – whether you're still respected and thought well of afterwards. As noted above, the reasons which successfully do this may change over time, whereas the notion of normative justification which philosophers of love are after ought perhaps to be more robust than this.

Although, it is true that social justification and normative justification sometimes overlap, in the sense that they identify the same reasons as good or bad reasons. Think about the example we gave earlier of saying that the reason for loving someone is that they have a lot of money. This would not be socially acceptable. But we can envision a society where this is in fact acceptable as a reason for love. Rationalists would not accept having money as the reason for loving someone in any circumstances, because that would make the beloved *fungible*, i.e. interchangeable for anyone who has money. For them, it could never be the case that having money is not inapt. Here, it suffices to illustrate how despite pointing to the same reason, the sense of social justification is contingent, and the normative justification philosophers of love discuss is not.[8]

However, this is to overlook that social justification is powerful despite being contingent because it is a way – and not the only way – that individuals *feel* justified in their attitudes and choices. Philosophers of love may be more interested in whether individuals simply *are* justified or not, regardless of feelings. In other words, there is a difference between a person being justified and a person feeling justified,

[8] Other philosophers of love are also already poking at this divide between explanation and justification as well. For example, Han (2021) states that philosophers of the reasons-view invoke that "love seems explained in virtue of being rationalized by reasons" (pg 5), and that "Even if normative reasons are distinct from explanatory reasons, there nevertheless tends to be an intimate connection between them" and that "it isn't as if there's no connection between these two sets of reasons. Some of the explanatory reasons for Andrew's coming to admire Candice, after all, are ones that relate him to the normative (fittingness) reasons for it" (pg 6).

and philosophers of love are concerned with the former. However, we think that these matters of *feeling* justified and *being* justified are not so easily kept separate. Firstly, the *felt* sense of justification involved in love is an important component to making it the recognisable phenomenon it is, and secondly, this *felt* sense of justification can lead to the relevant changes within a relationship that bring it to be normatively justified, in the way described by rationalists.

On Tiberius' view (2010), a *felt* sense of justification does a lot of work in explaining why our values and commitments in life are as powerful as they are and are recognised as such by ourselves – that is, we recognise them as meaningful because of the *felt* sense of justification they have for us. We have to *feel* that our values are justified in order to really take ownership of them and make sense of their influence in our lives, often structuring our entire lives. She says:

"If we focus on the value commitments that shape the decisions we make about how to live our lives, the need for justification is independently motivated. Failing to find any justification can undermine our commitments (a phenomenon that is often the source of the existential crises that befall people at a certain stage of life). The role that our value commitments play in our lives makes this sense of justification essential." (page 30).

Here, Tiberius helpfully highlights the close link between a felt sense of justification, and the requirement for identifying independent justification at all. That is, something feeling justified to us is an important component in recognising it as something for which there could be external standards of justification. Regardless of whether external standards of justification are met for something, (for instance, is 'love of nature' really a justified value to have?), it is something *feeling* justified which explains it having the powerful impact on individuals which it has. For instance, giving them and their lives a sense of meaning, dedicating a lot of time and effort to it, and accepting it as a core part of their identity. And she suggests that love is the same. To some extent, our love has to *feel* justified to us for us to recognise it as love at all (ibid, 28). She expands:

"Taking your commitments to be justified in the sense I intend is different from providing reasons for them. Taking yourself to be justified, in my view, means that you think a story could be told, not that you are actually prepared to tell it" […] "this story need not be one that is philosophically illuminating; nor does it need it be a story that speaks to universal reasons. Instead, the story might have a lot to do with how you feel when you are guided by these commitments" (page 28).

She emphasises that just because individuals might struggle to formulate reasons clearly and provide them, this doesn't mean that their values are not justified or that they do not feel that they are justified. They are just justified in a different sense - they have *felt* justification. We also take this talk of 'stories'

to be very easily construed as talk about confabulations. Although Tiberius here says that the story only 'could' be told and you may 'not be prepared to tell it', once you are pressed by the social world to provide reasons for your loving someone, you are likely to start putting this story together.

Notably, something 'feeling' justified for you is exactly the kind of attitude for which we are unlikely to have complete and accurate introspective access into the causes of[9], whether this be for a treasured value like a love of nature, or for loving your partner deeply. And here we see again the cognitive gap, or limited epistemic access, which agents are likely to fill with confabulation. But then, in turn, one's confabulations can further illuminate and entrench one's sense of these values and one's loving as being justified - because you have now formulated reasons for it, which you take to be good reasons (remember that confabulators are motivated to give reasons which justify these attitudes in a social sense, for themselves and others). In the next section, we explain how this dynamic interaction between confabulation and felt justification, can influence and shape one's relationship such that those reasons do indeed become normative reasons, and the love is justified.

Seeking *felt* justification explains some of the effects seen in empirical studies which include investigator manipulation in getting participants to introspect and give reasons for their loving someone. We described that optimism bias is one cause of distortions in the appraisals of one's partner and relationship, but in some studies investigators have directly introduced manipulation conditions which affected the appraisals given. In a study by Seligman et al (1980), participants were asked to give reasons why they were attracted to their partner, and about how committed they were to their partner. Participants who were prompted to give external/instrumental reasons for being with their partner, rather than being prompted to give intrinsic reasons for being with their partner, were more likely to rate their attitudes towards their partners more negatively, and more likely to predict not still being together in the future.

We suggest that individuals rate their love more positively when they were prompted to think of intrinsic reasons for this because intrinsic reasons provide them with much more *felt* justification. Individuals clearly do not feel that extrinsic reasons justify their love, and this leads them to rate their love more negatively. There is nice compatibility here between what individuals are clearly valuing as justifying their love, and the second sense of 'appropriate' which we outlined in section 1; that the love be rooted in the personal qualities of the beloved. We also see that individuals' attitudes change over time and in the light of having given reasons for their love (Wilson & Kraft 1993; Wilson et al 1984;

[9] This would be inline with psychological literature showing that introspection can be unreliable; see Lawlor 2003, Wilson 2002, Carruthers (2011) for discussion and Bortolotti (2009) for a helpful overview and analysis.

Wilson & Hodges 1993). In the next section, we elaborate even more on how these processes can take place, such that reasons which individuals give for their love become normative reasons because they do indeed make the love justified and appropriate.

4. From felt justification to normative reasons

In this section, we now turn to the rationalists and explain how these confabulatory reasons for love and the felt justification they bring, end up forming normative reasons for that love. We do this by highlighting the role that they play in a long-term, dynamic process of loving someone for certain reasons, and articulating those reasons. This deeper sense of one's love being justified, in the light of the reasons given in confabulations, can mean that the relationship then goes on to change and be shaped by those confabulatory reasons. Here, we describe some mechanisms for this.

4.1 Ill-grounded confabulations

When you formulate an answer to why you love someone, you have come to think about it more explicitly than you have done so before. Once you've put voice to that reason, you then start to notice it more in your daily life together afterwards. This is what happens to Harry in the example we offered in section 1. He had noticed before that Sally takes a long time to order a sandwich, but only when prompted – only when placed in the reason-giving stance – he picks that fact as one of the reasons for his love. We accepted that this may not in fact be a normative reason (we said that, if Sally were to start ordering sandwiches more quickly, he'd probably still love her). This means that, on our picture, this is an ill-grounded confabulation given by Harry about why he loves Sally. However, we claim now that it is possible that by formulating that reason explicitly, and bringing that reason into explicit deliberation or reason-giving (say, for example, in her wedding vows, or when reminiscing that decisive night telling their love story), Harry *does* come to love Sally for that reason, which was initially confabulated. Ordering a sandwich slowly may in fact come to be part of the actual set of reasons S whose absence would mean that Harry's love for Sally is unjustified.

Ordering a sandwich slowly is not a personal quality of Sally's, but it is nevertheless something that Harry *likes*. Sam Shpall (2018: 114), who considers liking a necessary element of love, states that "to like something is to be disposed to enjoy it, feel affection for it, experience attraction to it". It is easy to cite things that we like about others when we are in the reason-giving stance. What you like about

your partner, however, is unlikely to be firmly locked in place from the very beginning, but rather slowly evolve over time. By formulating explicitly the things that we like about the people we love, we are bringing these things to the forefront. In other words, by thinking and deliberating about the things that we like, even if it is through confabulation, we may cement that those are, indeed, the things that we like about that person. Before this provocation, we are likely to have had cognitive gaps there, for exactly why we have found ourselves liking someone.

As we explained in section 1, the first sense of appropriateness is that love is prudentially or morally valuable, and liking is an indisputable source of prudential value. By thinking frequently about what we like about the person – even if we confabulate – we cement what we like about the person, and thus the confabulated reason becomes a normative reason, i.e., becomes appropriate in the first sense. This is a mechanism which can unfold after having given ill-grounded reasons which are nonetheless true. For instance, it may be true that Sally orders slowly and that Harry liked Sally for her slow ordering, but this nonetheless was not something which normatively justified his loving her. (Remember that it did not meet the counterfactual condition. He would have continued loving her without it). However, now that he has thought about this feature more explicitly given his confabulating, it is something he is aware of, more aware of his fondness of, and the counterfactual condition comes to be met and his love for Sally is normatively justified.

4.2 False confabulations

We now turn to confabulations which are *false*, and argue that they can become normative reasons for love in that they can make the love appropriate in the second or third senses of the term. The second sense of justification was that reasons are apt (i.e. that the reasons given are intrinsic properties of the beloved) and the third sense was that reasons pick up real facts about the world. In our example, Zara says that she loves Chris because he is generous, but this is false - Chris is simply not generous.

Dean Cocking and Jeanette Kennett (1998) identify two necessary elements of love: direction and interpretation. Direction consists in changing the beloved's behaviour, while interpretation consists in changing the beloved's self-conception. They give the example of two friends, one of which loves ballet and the other who doesn't. Because the former friend cares about spending time together, they will start going to the ballet–their friends' interests will *direct* her actions. Further, by continuing going to the ballet with her friend, she may actually become interested in ballet, so ballet becomes a shared

interest. Secondly, when we care about people, we take to heart their judgement about who we are, that is, their *interpretation*. For example, Judy keeps telling his friend John that he always likes to be right and takes himself too seriously, a trait that John had not noticed about himself. John not only realises that he takes himself too seriously, but may actually change that trait in light of his relationship with Judy:

"Beyond making salient an existing trait of character, the close friend's interpretation of the character trait or foible can have an impact on how that trait continues to be realized. Within the friendship John's liking to be right may become a running joke which structures how the friends relate to each other. John continues to insist that he is right; however, his insistences are now for the most part treated lightheartedly and take on a self-consciously ironic tone. And John may be led by Judy's recognition and interpretation of his foibles to more generally take himself less seriously. Thus, John's character and his self-conception are also, in part, drawn, or shaped, by his friend's interpretations of him" (Cocking and Kennett 1998: 505).

Applying Cocking and Kennett's view to our example of Zara and Chris's example clarifies how confabulated reasons can become justificatory in the second and the third sense. The second sense was aptness – that love is grounded on intrinsic properties of the beloved. If Zara says that Chris is generous but this is not true, this interpretation might nevertheless make Chris change when seeing himself through Zara's eyes. Just like Judy's interpretation changes John, Chris may be changed by Zara. In Cocking and Kennett's example there is an implicit assumption that Judy is in fact picking accurate facts about the world, but there is nothing in the view that makes it a requisite. In the example, Zara's reason for loving Chris becomes appropriate by picking in a feature of Chris–being generous–that is in fact brought about by Zara's (false) confabulation. There is some empirical support that these kinds of changes can and do take place in relationships; Murray and colleagues (1996b) found that the idealisation of partners had self-fulfilling effects, in that partners came to embody those idealised images that the other had of them. They write, "Intimates who idealized one another appeared more prescient than blind, actually creating the relationships they wished for as romances progressed" (1996a, 1155).

However, another possibility is not that the facts change such that Chris becomes generous, but rather that once Zara formulates more explicitly the idea of loving Chris because he is generous, she really starts to notice all the ways that he falls short of this. If there are no prospects for Chris becoming more generous, and indeed he does not change, Zara could realise that Chris is not

generous and that therefore she does not have this reason to love him. In this case Zara still gains epistemically from confabulating, in realising that Chris is not generous, and there was some element of self-deception here. If Chris does change, however, then Zara's confabulation played a role in facts of matter changing such that what is confabulated about becomes true, those reasons become normative and the love is justified. This is because they now satisfy the second sense of appropriateness (Chris' generosity is one of his personal qualities), and the third sense (it picks up real facts about the world).

A final note on these mechanisms. When these mechanisms follow from confabulation, we do not envisage that agents simply get closer to the facts of the matter regarding what normative reasons justify their love. Or in other words, we do not envisage that agents are simply managing to find out what was in their 'cognitive gaps' all along. Rather, the very act of confabulating the reasons we do, shapes the normative reasons which are in play. Our confabulations, as a mode of enquiry, influence and shape what 'answer' is there to uncover. Confabulation as a mode of inquiry therefore interacts with, and is not completely separate to, the facts of the matter regarding normative reasons for love. We think that a mindshaping account of how we interpret ourselves and others can best capture this 'regulative' dimension of self and other-ascriptions, which include confabulated ascriptions (such as, for example, that Chris is generous) (McGeer 2007; 2015).

Furthermore, there will be a constant back-and-forth dynamic, where individuals have feelings about their relationships, and make sense of them with explicit appraisals and confabulations, which go to shape and deepend feelings and particularly feelings of justification, which go on to shape explicit confabulations, and so on. In essence, we do not see a clear line to draw between the shaping of one's love from the influence of confabulation (mechanisms described above), and of finally loving someone in a way which is normatively justified in the way rationalists describe.

In this section, we have demonstrated the relevance of confabulated reasons for rationalists. We've explained how confabulated reasons can indeed form the normative reasons for love which rationalists embrace, and this is importantly through how confabulated reasons are driven by and further shape one's *felt* sense of justification within a relationship. With this account in hand, we now address what the introduction of confabulation and felt justification for love brings to anti-rationalists and the wider debate on whether love can be justified.

5. Anti-rationalists and the wider debate

Our account has important ramifications for the wider debate on the philosophy of love. In section 1, we saw that Rationalists and Anti-Rationalists are in principle starkly divided with respect to reasons of love. By introducing the notion of felt justification, we have shown that Anti-Rationalists cannot escape justification wholesale.

Let us go back to the victim of physical abuse Smuts gives as an example, who is only able to give explanations about why she loves her abuser, which are not accepted by her concerned friends as acceptable justifications. Now, compare the woman in the example with the famous response given by Michel de Montaigne (1580 [2004]: 10) in his elegy for his friend Etienne de la Boetie: "If I am pressed to say why I loved him, I feel it can only be explained by replying: 'Because it was he; because it was me'". The reasons given by Montaigne and by the woman in Smuts's example are identical. Our account is able to explain the difference between the two examples. They differ insofar as the woman in the example, unlike Montaigne, is unable to feel justified given that her love is prudentially disvaluable—which in turn, obliterates the possibility of social justification. This comes at a very low cost to the Anti-Rationalists: they need not accept that love is the sort of phenomenon that can be epistemically justified in terms of reasons. They only need to accept that *offering* reasons for love can itself be constitutive of love, which is something that Frankfurt (2004: 67) already does: "Appreciating the value of an object is not an essential condition for loving it. It is certainly possible, of course, for judgments and perceptions of that sort to arouse love".

A staunch Anti-Rationalist may still argue that their view is incompatible with ours because of our emphasis on justification, which they reject. To respond to this objection, we briefly look at Frankfurt's view of love and how it relates to his view on the structure of the will. For Frankfurt, love is a form of concern, a 'volitional necessity' – meaning that our will is constrained in a way that does not allow us to *not* be concerned about the object of our love. However, elsewhere in his discussion on the structure of the will, Frankfurt gives great importance to the notion of wholeheartedness, which refers to an agent's endorsement to the expressions of their will. His best-known example is that of the Willing and the Unwilling addict (1971): both agents have their will constrained in a way that they cannot choose to not desire drugs, but they differ with respect to their higher-order desires. The Unwilling addict has a lower-order desire for drugs, but a higher-order desire to not desire drugs. Wholeheartedness necessitates an alignment between higher- and lower-order desires.

Now, going back to the notion of felt justification that we are endorsing here, let's think of two people, Ava and Allie. Ava knows she loves Allie and wants to be with her, but she does not know why (i.e., she has no epistemic access to her reasons for loving Allie). Allie knows she loves Ava and wants to be with

her, but she knows she shouldn't want to be with Ava because she is aggressive. When asked by her worried friends (like in Smuts's example), Allie will confabulate reasons. To simply say 'I don't know why I love her, although I know I shouldn't' would reveal an incongruence between her lower- and higher-order desires. If we ask Allie, she in fact is aware of this incongruence between her lower– and higher-order desires (she wants to be with Ava, but she does not want to want that).

Both women are confabulating. Ava will confabulate in the way we have shown that confabulation happens if we understand love as the Rationalists do, i.e. as an appraisal of the beloved's value. Her aim is to fill her cognitive gap regarding why she loves Allie, and thus feel justified. Importantly, Allie will confabulate in a different way. She will confabulate aiming at wholeheartedness: to align her lower and higher-order desires, or at least to show others that these are aligned. Allie wants to feel that she is not throwing her life away for no reason: she wants to feel justified as well.

Ava's example would not make sense from an Anti-Rationalist perspective, given that Anti-Rationalists would not accept that there is a cognitive gap at all: there is no gap because there are no normative reasons for her to be overlooking. But if we re-interpret our view in the way we just did with Allie's example, we can see that our view is at the very least *compatible* with the Anti-Rationalist chance. If they see wholeheartedness as a necessary element of a well-structured will, then they should acknowledge the importance for agents to feel justified about their commitments, as Tiberius argues.

Our account then erases any stark divide between Rationalists and Anti-Rationalists by offering a point in common—felt justification—that they have so far been blind to. In that sense, the view we set up here is reminiscent of the one put forward by Bennett Helm (2010), who in fact defines emotions as 'felt evaluations': feelings with positive or negative valence that are accompanied by intentional states. Helm similarly argues that the divide between Rationalists and Anti-Rationalists should be overcome, since in his view, love is both an appraisal and a bestowal of value. More precisely, for Helm love consists of concern for the beloved's identity, which is understood in terms of her values and is part of a wider pattern expressed with different person-focused emotions (like pride or shame, for example). We are sympathetic to the view, but the problem is that even if accurate, the account is too complex to be assessed in terms of the reason-giving stance. Helm gives the example of his spouse's valuing the activity of bagpiping. From the point of view of a straightforward quality theory, if he was placed in a reason-giving stance, he would need only to cite that value: "I love her because she is a bagpiper". This is because in most quality theories, the beloved's quality directly justifies love. Helm's picture is significantly more complex:

"[I]n being proud of my wife for winning the bagpipe competition, my pride is focused on her and subfocused on her playing bagpipes. To feel this pride is to be committed to the import she has as this person and so to the import bagpipes have as a part of her identity as such" (Helm 2010: 156).

It seems plausible that Helm would accept one of two views: either he would be a pessimist about epistemic access (like Naar) or he would accept a simplified version that is similar to a straightforward quality view ("I love my wife because she is a bagpiper"). If the latter, then we are illuminating an aspect–epistemic access–that he does not address; if the former, we have already argued that we think pessimism is unwarranted.

6. Is confabulation good for relationships?

We have spent a lot of time spelling out the *good* things which can come from giving confabulated reasons when in the reason-giving stance. Some may find this to be a very optimistic picture. Does this mean that confabulation in the contexts of love is a good thing? Can't confabulation contribute to the formation and maintenance of bad relationships?

We have explained in detail the mechanisms by which confabulation can contribute to the shaping of relationships in ways which make them justified, and turn those confabulated reasons into normative reasons for that love - because they come to be appropriate in at least one of the three senses of the term. But we don't argue that this is guaranteed, and in fact, we accept that confabulation can hide shortcomings of our relationships and partners to ourselves, or even - in virtue of their shaping powers - make relationships worse.

It could be the case, for instance, that Chris continues to be selfish, and Zara's confabulation means that she continues to live in blissful ignorance with regards to Chris' (lack of) generosity. Even more worryingly, Chris could come to be actively abusive, withholding money and necessities from Zara and becoming increasingly controlling. Zara's confabulations about Chris being generous might have not only stopped her from seeing that he isn't generous, but even contributed to him becoming abusive because of the sense of validation and justification they gave him with regards to his (very low) levels of generosity[10].

[10] See Jefferson 2020 for discussion of how confabulation and rationalisation can lead to a deterioration in one's moral character.

Another, worrying possibility is that confabulation plays a role in prolonging or exacerbating instances of love which strike us as unhealthy and very starkly at odds with facts of the matter. For instance, imagine someone who takes themselves to be deeply in love with a celebrity, where perhaps the lover even convinces themselves that the celebrity loves them back. There will be plenty of counter-evidence for these claims, but of course, confabulations are poorly rooted in evidence. In place of a poor understanding of one's own state of mind and the reasons for it, individuals might confabulate reasons that their beloved does indeed love them back, despite all evidence which point to the opposite.

We accept that these cases are all possible. However, we refine further what exactly confabulation is 'on the hook' for. We argue that there are two other factors which play important roles in cases where confabulation contributes negatively to relationships: the first is features of the relationship, and the second is the surrounding environment of the agent. We discuss each of these factors in turn.

In other work, one of the authors has argued that individuals having certain self-relational skills and attitudes makes the difference between whether confabulating can be part of the improvement of one's character, or just a chronic way of hiding one's shortcomings from oneself (Murphy-Hollies 2023). Here, we want to make some tentative first suggestions of features of a *relationship* (as opposed to an individual), which make it the case that that relationship can benefit from confabulation. For instance:

- Listening and communication: so that suggestions made by members of the relationship can be expressed, be listed to and be understood by the other, and finally be incorporated. This will no doubt involve long-term, consistent conversation and 'checking in' with regards to progress and understanding.
- Open-mindedness: this better enables the processes of direction and interpretation described by Cocking and Kennett (1998), to take place.
- Being optimistic about the relationship, about its resilience, longevity, and about your partner: it's going to be helpful to have *positive* conceptions of the relationship and of prospects for *meeting* those conceptions. This gives relationships the best chance of embodying them.
- Interest and Attentiveness: There won't be much point in voicing the idealised notions of one's relationship and one's partner in confabulations, if the other partner has no interest or pays no attention to these expressions. This lack of motivation means that positive changes are unlikely to happen.

- Patience: changes take time to implement. For instance, for new habits to become ingrained, to develop a taste for something new (such as the ballet), to become attuned to previously overlooked values (to take advantage of opportunities to be generous).

None of these features appear to be particularly novel or surprising as suggestions for what makes relationships go well. However, we explain that the reason why they make relationships go well is that *they are skills and attitudes which allow members of relationships to embody their idealised concepts of love and relationships.*

However, couldn't these very skills and attitudes be used to make relationships worse? Answering this brings us to the second factor we named above; one's environment. The skills above could indeed make a relationship harmful, *if the idealised concepts being used are bad.* In these cases, then these skills are only bringing a relationship closer in line with a bad/harmful idealised concept of the relationship. We accept that this is possible, if members have a totally skewed concept of what good or ideal relationships and love consists of. We suggest that this is most likely to stem from being in an oppressive society, when one is saturated in cultural messages which instil harmful notions of love.

Social negotiations tend to follow from confabulations. After all, this is one of the motivations of confabulating - to remain a part of the social world and be well-regarded by others (Bortolotti 2018a; Bergamaschi Ganapini 2020). But this social dimension is indeed a double edged sword. In the aftermath of Zara's confabulation that Chris is generous, her peers could well point out how unlikely this is. After all, Zara had a cognitive gap here - she cannot simply introspect further and realise that something else explains her love of Chris. Social interaction is something which allows us to fill these gaps (this is likely because others do not share your own motivations and biases to preserve your positive self-concepts). This can be fruitful and reflects the moral and epistemic benefits which the social world can bring.

However, this also depends on the type of social world Zara is submerged in. We can easily imagine Zara receiving very bad advice. Perhaps she says that Chris is generous, but her friends tell her that his behaviour displays weakness on his part and he's not being 'a real man', and that Zara should hand over all control and decision-making in the household to him. In these cases, the main problem is the misguidedness of her social environment, rather than confabulation itself being bad. And in these cases, we simply accept that confabulation can exacerbate issues which are rooted in non-ideal cultural and social environments. But this does not mean that we cannot accept that confabulation

can (and often does) play a role in the *improvement* of relationships, via the methods we describe in section 4.

This means that confabulation might be a part of some ugly pictures, but it doesn't do this alone, and doesn't introduce harm all by itself. Confabulation can also be a part of justified love, in becoming normative reasons for love.

Bibliography

Abramson, K., and A. Leite. (2011). Love as a Reactive Emotion. *The Philosophical Quarterly*, 61(245): 673–699.

Bergamaschi Ganapini, M. (2020). Confabulating reasons. *Topoi*, 39(1), pp.189-201.

Bortolotti, L. (2009). The epistemic benefits of reason giving. *Theory & Psychology*, 19(5), pp.624-645.

Bortolotti, L. and Cox, R.E. (2009). 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. *Consciousness and Cognition*, 18(4), pp.952-965.

Bortolotti, L. (2018a). Stranger than fiction: costs and benefits of everyday confabulation. *Review of philosophy and psychology*, 9(2), pp.227-249.

Bortolotti, L. (2018b). Optimism, agency, and success. *Ethical Theory and Moral Practice*, *21*(3), pp.521-535.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. OUP Oxford.

Delaney, N. (1996). Romantic Love and Loving Commitment: Articulating a Modern Ideal. *American Philosophical Quarterly* 33(4):339-356.

Cocking, D. and Kennett, J. (1998). Friendship and the self. *Ethics* 108 (3):502-527

Cocking, D., and Kennett, J. (2000). Friendship and moral danger. *The Journal of Philosophy* 97(5):278–296

Elder, A. (2014) Why Bad People Can't be Good Friends. *Ratio* 27: 84-99.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy* 68(1):5–20.

Frankfurt, H. (2004).*The Reasons of Love*. Princeton: Princeton University Press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4), p.814-834.

Han, Y. (2021). Do We Love For Reasons? *Philosophy and Phenomenological Research* 102: 107– 127.

Helm, B. (2010). *Love, Friendship, and the Self: Intimacy, Identification, and the Social Nature of Persons*. Oxford: Oxford University Press.

Isserow, J. (2018). On Having Bad Persons as Friends. *Philosophical Studies* 175 (12): 3099–116.

Jefferson, A. (2020). Confabulation, rationalisation and morality. *Topoi*, 39(1), pp.219-227.

Jeske, D., (2007). Friendship and Reasons of Intimacy. *Philosophy and Phenomenological Research*, 63(2), pp.329-346.

Jollimore, T. (2011). *Love's Vision*. Princeton, Princeton University Press

Keller, S. (2000). How do I love thee? Let me count the properties. *American Philosophical Quarterly* 37(2):163–173.

Kolodny, N. (2003). Love as Valuing a Relationship. *Philosophical Review*, 112(2): 135–189.

Kroeker, E. (2019). Reasons for Love. In Adrienne Martin (ed.), *Routledge Handbook of Love in Philosophy*. London: Routledge.

Lawlor, K. (2003). Elusive reasons: A problem for first-person authority. Philosophical Psychology, 16,549–564.

Mason, C. (2022) What's Bad about Friendship with Bad People? *Canadian Journal of Philosophy* 51 (7):523-534

McGeer, V. (2007). The regulative dimension of folk psychology. In *Folk psychology re-assessed* (pp. 137-156). Dordrecht: Springer Netherlands.

McGeer, V., (2015). Mind-making practices: The social infrastructure of self-knowing agency and responsibility. *Philosophical Explorations*, 18(2), pp.259-281.

Mercier, H. and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2), pp.57-74.

Montaigne, Michel de (1580 [2004]). *On Friendship*. London: Penguin Classics.

Murphy-Hollies, K. (2023). The Know-How of Virtue. *Journal of Applied Philosophy*.

Murray SL, Holmes JG, Griffin DW (1996a) The benefits of positive illusions: idealization and the construction of satisfaction in close relationships. *Journal of personality and social psychology*, 70(1):79–98.

Murray SL, Holmes JG, Griffin DW (1996b) The self-fulfilling nature of positive illusions in romantic relationships: love is not blind, but prescient. *Journal of personality and social psychology*, 71(6):1155–1180.

Naar, H. (2017). Subject-Relative Reasons for Love. *Ratio* 30(2): 197–214.

Naar, H. (2022). *The Rationality of Love*. Oxford: Oxford University Press.

Örulv, L. and Hydén, L.C., (2006). Confabulation: Sense-making, self-making and world-making in dementia. *Discourse Studies*, *8*(5), pp.647-673.

Pismenny, A. (2021). The Amorality of Romantic Love. In R. Fedock, M. Kühler y R. Rosenhagen (eds.), *Love, Justice, and Autonomy: Philosophical Perspectives*. New York: Routledge.

Protasi, S. (2016). Loving people for who they are (Even when they don't love you back). *European Journal of Philosophy* 24 (1): 214–234.

Seligman, C.,Fazio,R., & Zanna, M.(1980). Effects of salience of extrinsic rewards on liking and loving. *Journal of Personality and Social Psychology*, 38,453–460.

Setya, K. (2014). Love and the Value of a Life. *Philosophical Review* 123(3):251-280.

Stammers, S. (2020). Confabulation, explanation, and the pursuit of resonant meaning. *Topoi*, *39*(1), pp.177-187.

Shpall, S. (2018). A Tripartite Theory of Love. *Journal of Ethics and Social Philosophy* 13 (2):91-124.

Smuts, A. (2014), Normative Reasons for Love, Part II. *Philosophy Compass* 9: 518-526

Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence. *Consciousness and Cognition*, *33*, pp.548-560.

Tiberius, V, (2010). *The reflective life: Living wisely with our limits*. OUP Oxford.

Thomas, L. (1991). Reasons for loving. In: R.C. Solomon, and K.M. Higgins (eds.), *The philosophy of (erotic) love*. Lawrence: Kansas University Press.

Väyrynen, P. (2019). Normative Explanation and Justification. *Noûs* 55 (1):3-22

Velleman, J. D. (1998). Love as a Moral Emotion. *Ethics*, 109(2): 338-374

Wilson,T. (2002). Strangers to ourselves. Cambridge, MA:Harvard UniversityPress.

Wilson, T.D., Dunn, D.S., Bybee, J.A., Hyman, D.B. and Rotondo, J.A. (1984). Effects of analyzing reasons on attitude–behavior consistency. *Journal of Personality and Social Psychology*, 47(1), p.5.

Hodges, S.D. and Wilson, T.D. (1993). Effects of analyzing reasons on attitude change: The moderating role of attitude accessibility. *Social Cognition*, 11(4), pp.353-366.

Wilson, T.D. and Kraft, D. (1993). Why do I love thee?: Effects of repeated introspections about a dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin*, 19(4), pp.409-418.

Zangwill, N. (2013). Love: Gloriously amoral and arational. *Philosophical Explorations* 16(3):298–314.

Appendix

Paper 1: Why We Should be Curious about Each Other

Bortolotti, L. and Murphy-Hollies, K., (2023), 'Why We Should Be Curious about Each Other', in Philosophies, 8(4), p.71. DOI: https://doi.org/10.3390/philosophies8040071

philosophies

MDPI

Article

Why We Should Be Curious about Each Other

Lisa Bortolotti * and Kathleen Murphy-Hollies

Philosophy Department, University of Birmingham, Edgbaston B15 2TT, UK; klm276@student.bham.ac.uk
* Correspondence: l.bortolotti@bham.ac.uk

Abstract: Is curiosity a virtue or a vice? Curiosity, as a disposition to attain new, worthwhile information, can manifest as an epistemic virtue. When the disposition to attain new information is not manifested virtuously, this is either because the agent lacks the appropriate motivation to attain the information or because the agent has poor judgement, seeking information that is not worthwhile or seeking information by inappropriate means. In the right circumstances, curiosity contributes to the agent's excellence in character: it is appropriate to praise the agent for being curious, blame the agent for not being curious, and also prompt the agent to cultivate such curiosity, at least in some of the relevant contexts. We believe curiosity can also manifest as a moral virtue when it helps an interpreter view a speaker as an agent with a valuable perspective on the world. Especially in interactions where either there is a marked power imbalance between interpreter and speaker, or interpreter and speaker have identity beliefs that lead them to radically different worldviews, curiosity can help foster mutual understanding, and prevent the interpreter from dismissing, marginalizing, or pathologizing the speaker's perspective.

Keywords: curiosity; agential stance; epistemic virtue; moral virtue; mutual understanding

Citation: Bortolotti, L.; Murphy-Hollies, K. Why We Should Be Curious about Each Other. *Philosophies* **2023**, *8*, 71. https://doi.org/10.3390/philosophies8040071

Academic Editor: Genia Schönbaumsfeld

Received: 1 June 2023
Revised: 7 July 2023
Accepted: 20 July 2023
Published: 4 August 2023

1. What Is Curiosity?

Some philosophers think of curiosity as an emotion because it is manifested in feelings, physiological arousal, and facial expressions, and can sustain the motivation to act in a certain way [1]. Specifically, curiosity leads people to fill newly discovered gaps in information, and thus it is thought to support learning and encourage the exploration of new sources. For other philosophers, curiosity is predominantly a desire, and has been defined as "a drive to know" [2] or a desire for the acquisition of new worthwhile information [3]. Some authors have already argued that curiosity can be an epistemic and moral virtue [4–6].

In this paper, following Elias Baumgarten [7], we mostly view curiosity as a disposition to attain new information that can manifest in *epistemically* virtuous behaviors. We also suggest that curiosity plays a special role in our mutual interactions, and thus can manifest in *morally* virtuous behaviors. As Baumgartner observes, the word "curiosity" comes from the Latin *cura* which can be translated as "care" or "concern" [7]. The word's etymology is relevant to the notion of curiosity that we aim to develop here: when curiosity is directed towards other people's experiences, it manifests as a form of caring. In a social exchange, an interpreter adopts the *agential stance* towards a speaker when the interpreter sees the speaker as an agent who has a valuable perspective on the world [8]. When the speaker's report is likely to be dismissed by the interpreter because the speaker has less authority or credibility than the interpreter, the interpreter's curiosity about the speaker can prevent the interpreter from dismissing the speaker's report. When the speaker reports a view that is not shared by the interpreter, the interpreter's curiosity about how the speaker arrived at that view may not resolve the disagreement but can lead to a further exchange of information resulting in enhanced mutual understanding. This in turn makes it less likely that the speaker's perspective is marginalized or pathologized.

Here is a plan of the paper. In Section 2, we consider how we find both negative and positive evaluations of curiosity in the philosophical literature [9]. Curiosity has been described as a vice and as a virtue. A curious attitude may take people too far in the pursuit of new information, making them disregard constraints and rules, disobey, and transgress. Eve in the Garden of Eden is thought to exemplify these costs of curiosity at the extreme. More contemporary versions of this idea are ubiquitous in novels and films, where an act of *hubris* is performed by mad scientists and lone heroes who fail to recognize their personal limits and end up having to pay a high price for their thirst for knowledge. Curiosity as a virtue is associated with a love of knowledge and a desire for inquiry, and with the explorations of innovators.

In Section 3 we ask what conditions need to be met so that the disposition to acquire new information can count as a virtue. The distinctive character of the curious person is valuing knowledge in its own right, without specific concerns with what practical advantages can be derived from that knowledge. That is probably why curiosity is often an attribute of scholars. However, curiosity is also associated with the frivolity and superficiality of gossip or the desire to know about other people's experiences for the purpose of judging, alienating, or excluding them. That is why having a curious disposition can manifest as a virtue, but this is not always the case: for curiosity to manifest as a virtue, the sought information must be of the type that is worth attaining, the motivation for attaining it must be a noble one, and the agent must exercise good judgement in pursuing new information.

In Section 4, we argue that curiosity can manifest as an epistemic and moral virtue in social interactions where mutual understanding is threatened. To support our point, we offer two examples: (1) interactions in clinical encounters characterized by strong power imbalances, where the patient is at risk of being objectified and silenced; and (2) conversations between agents who have apparently irreconcilable identity beliefs and different values, where the opponent's views are at risk of being pathologized or marginalized because there is no identifiable common ground.

2. A Brief History of Curiosity

In an article defending the view that Nietzsche was a virtue epistemologist, Mark Alfano quotes Nietzsche as saying that curiosity is the most agreeable of all vices. The form of curiosity Nietzsche considered a virtue was "an insatiable desire to solve novel, difficult problems and puzzles, and to discover or invent them when none are ready to hand" [10]. Nietzsche's enthusiasm for curiosity is especially interesting for Alfano, because curiosity has not otherwise had a very good press in the philosophical literature.

Alfano notices that curiosity is often presented as a vice: in particular, Christian thinkers see it as the trait responsible for Adam and Eve's original sin. It is curiosity that is responsible for their eating from the tree of knowledge and falling from heaven. This is a recurrent theme, that curiosity leads people to attaining knowledge or wisdom that they should not aim to attain. By indulging in their curiosity, people fail to recognize their own limits [11]. When we read the Bible the story of how Eve comes to eat the prohibited fruit of the tree of knowledge, we realize that her motivation for doing so may be aptly described in terms of curiosity:

> "Now the serpent was more crafty than any of the wild animals the Lord God had made. He said to the woman, 'Did God really say, "You must not eat from any tree in the garden"?' The woman said to the serpent, 'We may eat fruit from the trees in the garden', but God did say, 'You must not eat fruit from the tree that is in the middle of the garden, and you must not touch it, or you will die'. 'You will not certainly die', the serpent said to the woman. 'For God knows that when you eat from it *your eyes will be opened*, and you will be like God, *knowing good and evil*'. When the woman saw that the fruit of the tree was good for food and pleasing to the eye, *and also desirable for gaining wisdom*, she took some and ate it. She also gave some to her husband, who was with her, and he ate it. Then,

> the eyes of both of them were opened, and they realised they were naked; so they sewed fig leaves together and made coverings for themselves." (Genesis 3:1–7 [12], our emphasis).

When the effects of eating the fruit are presented as "having one's eyes opened", "knowing good and evil", and "gaining wisdom", it is not surprising that Eve's drive to know is what makes eating the fruit appealing. However, eating the fruit leads to the "fall of the whole human race" [13]. Christian writers such as Augustine [14] and Aquinas [15] consistently describe curiosity as dangerous, sinful, and vain, associating it with pride. In subsequent critiques of curiosity, up to the seventeenth century, the idea that being curious leads to people becoming proud—due to their acquired knowledge—is ubiquitous. And much more recently, in the growing literature on the attitudes that are responsible for people endorsing conspiracy theories, one contender is the tendency that some people have to "do their own research" rather than follow expert opinion [16]. This attitude seems to combine curiosity with epistemic arrogance.

A change of attitude towards curiosity, from a form of hubris to a commendable desire for learning, is advocated by Francis Bacon, who defends in his writings the usefulness of knowledge [17]. However, even Bacon insists on the distinction between knowledge sought *for vain curiosity*, which is likely to give rise to pride, and knowledge sought *for the purposes of charity and philanthropy*, which is an appropriate and morally laudable aim. A wholehearted acceptance of curiosity comes from Thomas Hobbes who describes curiosity as a *morally neutral appetite for knowledge*, suggesting that it is one of the characteristics distinguishing humans from nonhuman animals [18]. Gradually, with the Enlightenment, curiosity gains some respectability as a foundation for science and, in the current philosophical literature, it is frequently presented as an epistemic virtue.

What does it mean that curiosity is a *virtue*? To consider curiosity a virtue means that it contributes to excellence in character, and that, in at least some of the relevant contexts, it is appropriate to praise agents for being curious, blame them for lacking curiosity, and also encourage them to cultivate curiosity. This is what underlies the focus on curiosity in schools and academic institutions, where educators are meant to "facilitate, encourage and nurture" the motivation to acquire epistemic goods in their students [6] and universities often present themselves as "champions of curiosity" [9]. One concern is that in educational settings curiosity is superficially endorsed and not critically examined.

What is it for curiosity to be an *epistemic* virtue? All epistemic virtues are aimed at "improving epistemic standing" in terms of enhancing the person's own or other people's understanding or knowledge [6]. But each epistemic virtue contributes to this general aim in a distinctive way. Curious agents are motivated to acquire new epistemic goods if they believe that such goods are worthwhile. Often the motivation to acquire epistemic goods is demonstrated by the person's willingness to pay some cost for the epistemic goods to be acquired. Indeed, in our example of curiosity from the *Genesis*, Eve paid a high cost for the knowledge of good and evil that she attained by eating the prohibited fruit. She had to leave the Garden of Eden with Adam, and was punished by God, together with all women after her, by experiencing pain in childbirth.

The interesting thing about curiosity, in our view, is how it straddles being an epistemic virtue and a moral virtue. Curiosity is characterized by an interest in epistemic goods. When these goods are information about other *people* and their *perspectives*, though, curiosity is instrumental to people's capacity to build connections with others and empathize with the experiences of others, pursuing a socially flourishing life based on genuine regard for others. Curiosity can therefore contribute to the excellence of both moral and intellectual character [19]. Curiosity can be thought of as a disposition to think, feel, and act in ways which are conducive to gaining epistemic goods, including information about other people's perspectives that leads to enhanced understanding. This in turn can improve relationships. We hope to demonstrate how transformative this can be, when done well, in Section 4.

3. Is Curiosity Ever a Vice?

Thinking of curiosity as a virtue invites us to consider the Aristotelian idea of what the extremes of *excess* and *deficiency* could be which flank each side of the trait, representing vices. When the desire for knowledge is *deficient*, we get an intellectually lazy and small-minded person. When the desire for knowledge is *excessive*, we get a nosy person who indulges in gossip, or a mad scientist who puts others at risk in the name of extending their knowledge. Although some argue that curiosity can be a virtue [20] or a vice [3], it is not clear to us that curiosity is ever a vice. A closer look at the various dimensions of epistemic virtues and vices can help us appreciate that the traits associated with curiosity can fail to be virtuous without being vicious.

Jason Baehr argues that intellectual virtues are strengths of character that have something like truth, knowledge, understanding, and wisdom as their aim, and contribute to a person's worth [19]. He also identifies four dimensions of virtues (competence, motivation, judgement, and affect) which need to be in place for a trait to count as virtuous. Baehr considers the possibility that agents have an intellectual vice only if they are defective with respect to all four dimensions of an intellectual virtue. But, in the end, he rejects this picture because he finds that there are certain deficiencies in motivation and judgement that are more important than others.

In terms of *competence*, skill underlies a virtuous trait. Baehr suggests that, in the case of curiosity, the competent person is able to ask the right questions [19]. In the interpersonal context, this is the perhaps the first barrier which faced by the would-be curious interpreter; lacking the skills to know which questions to ask, and how to ask them, in order to gain worthwhile and insightful information about the speaker. Lacking competence may be a reason for an agent to fail to have the virtue of curiosity without necessarily having a vice, and in particular without having a vice of curiosity.

Second, the virtuous agent needs *appropriate motivation*. If agents wanted to learn about the world and other people's perspectives just to be praised and admired for their knowledge or to manipulate other people's behavior for personal gain, they would not have the virtue of curiosity. The virtue of curiosity requires that agents attain knowledge (or some other epistemic good) due to their love for knowledge. Interpreters who want to find out about the speaker's perspective to undermine it or ridicule it do exemplify a vice. But the vice is something other than curiosity and rather a kind of malevolence—a use of newly acquired knowledge for the purpose of harming others. Inappropriate motivation of this sort can give rise to what seems an excess of curiosity; being nosy and gossiping, for instance. Small-mindedness may give rise to a deficiency of curiosity in the passive and dismissive agent. These inappropriate motivations may lead to vicious behaviors but the vice does not lie in the trait of curiosity.

The third dimension of curiosity is *judgement*: agents need to exercise practical wisdom in their pursuit. The need for good judgement and practical wisdom in the exercise of virtue is exemplified by the fact that we do not attribute the virtue of courage to a person perpetrating a terrorist attack or the virtue of curiosity to a person who follows us home to discover where we live. In the case of curiosity, exercising judgement may be a matter of identifying a deserving epistemic goal, knowing when it is appropriate for the agent to pursue that epistemic goal, and establishing to what extent the goal should be pursued. Agents' lack of judgement can be an obstacle to their desire for knowledge being virtuous, but does not always mean that they have a vice. We shall return to this point shortly.

Finally, we come to the *affective dimension* of virtues, meaning that the virtuous agent needs to take pleasure in the pursuit of the virtue. So, the curious agent derives pleasure and satisfaction from the exercise of the relevant competence and from gaining the desired epistemic goods. The absence of these feelings of pleasure and satisfaction seems to suggest that the person's trait is not a virtue, but does not amount to a distinct vice. However, Baehr argues that willful ignorance, a lack of interest in the relevant competence, or a twisted pleasure in disregarding epistemic goods may amount to a vice if the agent is responsible for bringing about or preserving those attitudes. In essence though, lack of

affect and competence merely suggests that the disposition to know is not a virtue, whereas deficiencies in motivation and judgement more easily give rise to vices, though it is not the disposition to know itself to count as a vice in those cases.

An example can help illustrate our suggestion that a desire for knowledge can be less than virtuous without being a vice. There is some debate over how acceptable it is to ask people where they are from if there is some indication that they are foreign—where their being foreign may be indicated by their race, accent, cultural references, and so on. On some accounts, the question is motivated by the desire to know more about that person, a sort of healthy and friendly curiosity. On other accounts, the question can be seen as alienating and intrusive, potentially aimed at excluding and objectifying the other, and thus an instance where the desire to know ought to be reined in.

The agent who desires knowledge and has competence, appropriate motivation, and good judgement should be able to ascertain whether asking the question is the right thing to do, depending on the context and the nature of the interaction. Is the other person a stranger or a good friend? Are their origins something they are likely to want to talk about? Is the question something they might often be asked? The agent who is curious in a virtuous way should be able to draw on their previous experience and their current knowledge of the situation in order to inquire about other people's life history and culture with sensitivity.

In order to have the virtue of curiosity, our curious agent needs to be skilled in asking questions, derive pleasure from attaining the relevant knowledge, exercise good judgement, and be appropriately motivated to gain knowledge—there is no intention to alienate or exclude. If the agent asks someone where they come from without carefully considering the implications of the question, then we may wonder whether the agent lacks the virtue of curiosity or is curious in a vicious way. There are a few possibilities to consider here, which affect whether there is merely an absence of virtue or a vice. There may be something naïve and immature about the questioner who isn't picking up on cues about the best way to express their desire for knowledge. If the questioner continues over time to ask about provenience in an insensitive way, this entrenched insensitivity could signal that there is a systemic failure somewhere, perhaps in judgement or motivation. At best the questioner fails to be curious in a virtuous way, and at worst the behavior reflects another vice—such as malevolence. This does not seem to make *curiosity as such* a vice, but rather points to a different vice which looks superficially like or feeds off curiosity.

However, another option is that the virtue of curiosity is simply underdeveloped at that time, and so the insensitive questioning is not necessarily entrenched. Perhaps the agent has been lacking in opportunities to practice respectful questioning exchanges up to this point. Think about the child who learns more deeply how to manifest proper generosity, acting with a genuine and more mindful concern for the well-being of the other person, rather than from a well-intentioned keenness 'to help' which is nevertheless overbearing and unhelpful. Similarly, our questioner needs to learn to exercise curiosity whilst keeping concern for the other's wellbeing still very much at the heart of the exchange, mindfully navigating options and using good judgement. This agent may be said to have some degree of curiosity, but not the virtue of curiosity, because the disposition needs to be exercised with better judgement and more practical wisdom. When it comes to motivation, when pursuing valuable, personal knowledge about another person, the questioner should be mindful of the other agent's wellbeing. Note that either of these deficiencies of judgement and motivation could affect the *reliability* of the curious disposition, which determines whether the disposition attains the status of a virtue. Those lacking in wisdom and mindful motivation are likely to overlook opportunities to gain worthwhile knowledge. However, unless malevolence is involved, there is no vice.

So, in this picture of curiosity as a moral and epistemic virtue, curiosity encompasses the agent's skill for asking worthwhile questions, the agent's motivation to gain epistemic goods for their own sakes whilst respecting other agents, the agent's pleasure in the pursuit of knowledge, and the agent's good judgement in exactly how to go about pursuing

knowledge (e.g., what questions to ask or how to ask questions). If one of these elements were lacking, the virtue of curiosity would be absent, but not all deviations lead to the same outcomes. If the agent has a poor motivation for pursuing knowledge, then they might have some vice which we argue is not curiosity (such as malevolence). Whereas it is easy to identify the right motivation, as the agent must want attain knowledge for knowledge's sake, it is difficult to establish how reliable the disposition must be, and how wisely the agent must manifest it, for the drive to know to qualify as the virtue of curiosity. If our questioner is skilled, well-motivated, and enjoys gaining epistemic goods, but lacks good judgement and is insensitive, their bad judgement and insensitivity may either be a sign that there is some poorly discerned vice present or that the disposition to know need to be practiced further. In the latter case, the agent needs to learn how to express their curiosity with good judgement, so that the case of potentially insensitive questions such as "Where do you come from?" is navigated wisely. Unless there is some vice inhibiting the development of the disposition to know, temporary insensitivity amounts to a failure to have the virtue of curiosity rather than a vice.

In what follows, we want to offer two illustrations of how curiosity can have distinct advantages in social interactions and exchanges characterized by epistemic challenges. In both cases, curiosity as a disposition to attain new worthwhile information is manifested virtuously in the agential stance. What do we mean by "is manifested"? Let us use an analogy. Just as the disposition to give is manifested virtuously in donating money to a charity, so the disposition to attain new, worthwhile information is manifested virtuously in inquiring about another person's perspective on the world. Making a donation to a charity is an expression of generosity and adopting the agential stance is an expression of curiosity. Clearly, the disposition is not sufficient for the action. It is not sufficient to be generous for making a donation. Availability of funds to donate is also necessary for the donation to take place. Similarly, it is not enough to be curious to value the other person as an agent with a valuable perspective on the world. Adopting the agential stance also requires a recognition that the other can form their own perspective.

The disposition is not even necessary for the action. A wealthy politician can donate money to charity to gain popularity and a teacher can value a pupil's view on the world without being motivated to learn how the pupil arrived at that view. However, when there is a donation motivated by generosity, the donation expresses generosity. And when there is a successful interaction between people who are motivated to know each other's perspective on the world and view each other as agents, their attitudes express curiosity.

So, adopting the agential stance is an epistemically and morally virtuous manifestation of the disposition of being curious. Here, curiosity has benefits that are epistemic in so far as an interpreter's curious disposition furthers an understanding that could not be attained if the speaker's perspective were unheard, dismissed, marginalized, or pathologized. But the curious disposition also carries distinctive benefits that are moral in so far as it promotes better social interactions, avoiding the negative effects of exclusion and polarization.

4. The Agential Stance as an Expression of Curiosity

Baumgartner argues that curiosity translates into a special kind of caring that people demonstrate for each other when they are not necessarily in a close relationship [7]. His example is that of a teacher who notices how a pupil reacts towards a certain topic discussed in class, taking the time to ask 'Why?'. Baumgartner suggests that curiosity sustains a form of engagement which has both knowledge and caring as its positive outcomes, vindicating curiosity as an epistemic and moral virtue. Here, we would like to offer two cases where curiosity is manifested in the agential stance.

4.1. Curiosity against Silencing: "What Is Your Perspective?"

The first case we consider is an interaction between interpreter and speaker where there is a strong power imbalance: the speaker seeks support, and the interpreter has the means to provide that support via their status, role, experience, or expertise (child/parent;

student/teacher; patient/doctor; etc.). We argue that it is an epistemic and moral benefit if both parties, but especially the party invested with more power, can feel and express curiosity towards the other party's perspective, enabling a genuine exchange of views.

One type of encounter that is characterized by a significant power differential between parties is the *clinical encounter*. On the one side, there is a person who is potentially vulnerable and needs support (patient) and on the other side, there is a person in a position of authority who can offer, or is the gateway to, the required support (healthcare practitioner). Typically, in these encounters, the patient presents with a problem and has, at best, experience of how the problem affects themselves, whereas the healthcare practitioner is an expert, thanks to their training and their clinical experience, in identifying the causes of the problem and proposing solutions.

Exchanges between patients and practitioners can go well or badly. When encounters go well, they enhance trust and are conducive to patients seeking help in the future. When they go badly, they undermine trust. This may result into patients failing to comply with the suggested treatment or avoiding healthcare services when a new crisis occurs. In what we may want to describe as *bad* conversations between practitioners and patients, the practitioner does not solicit detailed descriptions of the problem and does not fully engage with the patient's report of the situation. For instance, in some mental health encounters, clinicians may not look at the patient directly, they provide minimal verbal feedback, and avoid answering the patient's direct questions: "When patients attempted to present their psychotic symptoms as a topic of conversation, the doctors hesitated and avoided answering the patients' questions, indicating reluctance to engage with these concerns" [21]. This reluctance may have many reasons, and some may be justified in a particular clinical context, but it communicates to the patient that the practitioner does not want to engage with the patient's perspective. In studies on patient experience, disengagement is often described as a lack of curiosity, and lived experience advisors explicitly mention the practitioner's curiosity as something that would dramatically improve patient satisfaction. Further, patients would be prepared to disclose more useful information about their condition if they had reasons to believe that practitioners were interested in hearing more.

In what we may consider as *good* conversations between practitioners and patients, the practitioner shows interest in the person's experiences and asks more details about such experiences in order to understand the symptoms better and have a better sense of what can benefit the patient moving forward: "Psychiatrists commonly [. . .] explored the impact of the symptoms on the patients' behaviour and functioning, and also questioned the validity of the beliefs by directly challenging them or offering alternative explanations" [22]. It is interesting that in this passage, the curious attitude of the practitioner is described as a form of *exploration* and it does not imply acceptance. The curious practitioner engages with the patient's reports but need not agree with the patient's description of their health journey. Active engagement starts with listening and exploring further but can also manifest as "directly challenging" or "offering alternative explanations".

Thus, not the only difference, but one very important difference between bad and good conversations is whether the practitioner is curious about the patient's experiences and communicates such curiosity. Is the practitioner's curiosity clearly expressed in the exchange with the patient? The practitioner's interest can become apparent when attentive listening and open questioning are used. Conversations flow well and the patient is open to disclose more when they feel heard and understood and detect engagement in the practitioner's verbal and non-verbal behavior. However, conversations fail to flow and are full of pauses and silences when practitioners do not show appropriate motivation, and act as if their only goal is to go through a box-ticking exercise for the purposes of identifying possible treatment options or completing a risk assessment.

As we saw, curiosity as an epistemic virtue involves having the appropriate motivation to seek new information. It also involves good judgement: being motivated to seek valuable information is not sufficient. The agent needs to know when and how it is appropriate to

pursue this valuable information. Showing empathy and interest in what the patient has to say encourages the patient to share their own experiences and makes the patient feel that they are right in seeking support from services. This enables both practitioner and patient to have a better exchange and pursue the epistemic goods their exchange can offer.

But here, curiosity has also moral pay-offs: when the patient feels heard and understood, their trust in the practitioner's willingness and capacity to help increases, with positive outcomes in terms of the quality of the therapeutic relationship and the patient's future engagement with health services. The moral value of curiosity here is that, by demonstrating their engagement, practitioners manifest their commitment to viewing patients not just as problems to solve but as agents with a valuable perspective to share.

In a recent project on how to preserve agency in young people using mental health emergency services, video-recorded clinical encounters were analyzed by experts in philosophy, psychology, psychiatry, and also young people with experience of mental healthcare services. In the end, the recommendation was that the practitioner should commit to seeing the patient as an agent [8]. During the successful encounter, the practitioner sees the patient as a subject of experience with a perspective on the world, acknowledging that such a perspective matters. The practitioner also recognizes that the patient's concern or request for support is legitimate and deserves attention: the patient's concerns are not dismissed but addressed. For the practitioner, the patient is not just an object of study or a unidimensional problem to solve, but a person with a variety of needs and interests that need to be taken into account in decision making. Further, the practitioner affirms the patient's capacity to contribute to positive change and does not exclude the patient from the process of decision making, leading to potential treatment options. In practice, this might mean that the practitioner is willing to explain the advantages and disadvantages of the available treatment options to the patient and elicit the patient's views about such options.

Genuine curiosity accompanied by attentive and empathetic listening underpins all the communication techniques contributing to a successful encounter, including the validation of the person's experiences, the legitimization of the person's concerns, the lack of objectification of the person, the affirmation of the person's capacity to contribute to positive change, and the person's involvement in decision making. Validation, which is the first step towards the adoption of the agential stance, can be practically manifested by asking open questions motivated by genuine curiosity about the person's perspective.

4.2. Curiosity against Polarization: "What Is Your Story?"

The second case we explore is an interaction where there is substantial difference in worldviews between interpreter and speaker, and the exchange risks becoming unproductive and polarized—this is sometimes characterized as *deep disagreement*. It is both an epistemic and moral benefit if each party expresses curiosity about how the other party arrived at their view of the world because the ensuing understanding can help build bridges between the radically different views and increase mutual understanding. Ultimately, mutual understanding may prevent an impasse where the original views become radicalized and more ingrained, and the exchange has no positive outcome.

When people hold different values and express incompatible identity beliefs, the exchange can become difficult, compromising the possibility of a productive dialogue and resulting in the speaker's views becoming even more ingrained and irreconcilable with those of the interpreter. In these situations, merely offering positive arguments for one's position and raising objections against the opponent's position do not help overcome the impasse. The cycle of statement, objection, and response can be met by rigidity and defensiveness and is unlikely to encourage opponents to consider the merits of alternative positions. Moreover, the effectiveness of that cycle is hostage to the assumption that interpreter and speaker share the same values and the same methodological assumptions about what counts as good evidence for the positions in the debate. However, this is not always the case. For a cherished position to be given up, that is, a position recognized as

central to the person's own system of beliefs, the very idea of what counts as good evidence against that position needs to be revisited as well.

Take climate change denialism as an example. Claire has been a denialist for ages. In order for her to admit that human actions have played a role in climate change, on the basis of the evidence from climate science presented by her opponent, Claire would have to believe that climate scientists can be trusted. However, denialism is often based on the conviction that the official source of information (in this case, climate science) is not to be trusted due to its being unreliable and corrupt. As a result, the data provided by such a source are downgraded by Claire to non-evidence. A denialist like Claire would not accept the latest climate science projections as evidence against her position, unless something caused significant parts of her belief system to change, including those relevant to determining which type of information counts as genuine evidence in the debate [23]. For Claire's opponent's argument to be persuasive, the scientific data earlier dismissed by Claire as a fabrication or as a piece of propaganda would have to be upgraded to evidence, evidence sufficient to accept the claim that humans are partly responsible for climate change. As previously mentioned, positions on climate change and other topical and political issues may be entangled with people's values and identities, so other changes in Claire's worldview may be necessary as well for her to be disposed to accept the opponent's position.

What does it mean that some beliefs are *identity beliefs*? Identity encompasses those aspects of a person that acquire a special significance because the person takes pride in them or feels that they are central to how they are or want to be [24]. These aspects are relatively stable. Entrenched positions people have on religion, politics, morality, and so on, are shaped by and reflect their values and commitments. Take attitudes to COVID-19 vaccine mandates as an example. Refusing the COVID-19 vaccine and viewing vaccine mandates as part of a conspiracy will make sense to a person who already mistrusts the government and pharmaceutical companies and generally opposes invasive medical interventions. Suppose Omari has refused the vaccine and protested against mandates during the pandemic on the basis of the belief that vaccines are ineffective and unsafe. Omari also suspects that the government and the healthcare authorities must have some ulterior motive to push vaccination. In order for Omari to change his mind and accept the offer of a vaccine, it is not enough that he talks to someone who cites a science paper or a reputable source confirming the protective role of the COVID-19 vaccine and denying its risks. That is because Omari's attitude towards the vaccine is only a small part of his general outlook on life. In order for him to change his mind and agree that getting the vaccine is a good idea, Omari would have to change the way he sees and presents himself. From a vaccine sceptic he needs to become a person who is open minded about vaccines.

So, what can people do when an exchange of arguments fails to bring about progress in a debate? A proposal is that interpreter and speaker need to find a common ground, even when they start from different views and values. Such common ground is more likely to be something that involves *how they feel* and *what they care about* than what they have arguments for. In *How Minds Change* David McRaney offers an overview of the psychological science aimed at disclosing how people give up beliefs that were very important to them for an extended period of time, beliefs that capture some of their values, are laden with emotions, determine their behavior, and are relatively stable. McRaney is interested in the theory behind the "conversions" that people experience when they leave a cult or abandon a conspiracy theory [25]. However, he is also asking whether the theory can explain the success of some techniques that have been used to get people to distance themselves from their ingrained beliefs: *deep canvassing* and *street epistemology* among others.

Deep canvassing is a technique aimed at identifying some standards for respectful conversations with people who report a fixed position. The main recommendation is for interpreters to learn about the experiences surrounding the content of the speakers' position: the conversation becomes an exchange of emotionally and personally significant events that are explored together and that reveal the values of interpreter and speaker. In

the conversation, then, there is no longer an exchange of arguments, but an exchange of *stories*. Participants talk about how they came to their position on the relevant issue and why that position has now become so important to them. By sharing the reasons why that particular position occupies such a special place in their worldview and self-conception, the opportunity arises to extend the common ground between participants in the exchange.

Whereas the exchange of arguments is finalized to persuade the opponent that they are wrong, the exchange of stories is primarily aimed at making better sense of the other person's perspective. Asking about the other person's experiences and personally significant stories is again an expression of the interpreter's genuine curiosity about the speaker. This curiosity motivates the interpreter to inquire about the speaker's view and how that view was formed. It also motivates an inquiry into the emotional significance of that perspective. Virtuous curiosity here is not just a disposition to attain information about the speaker. It is important to exercise good judgement and practical wisdom in order to avoid insensitive questioning, as discussed earlier with respect to the question where someone is from. If the interpreter is genuinely interested in knowing how the speaker formed beliefs and opinions and is also skilled and wise in questioning, valuable information will be sought without intrusiveness. This may require the interpreter giving some gentle pushbacks without judging or alienating the speaker, and raising the possibility of different points of view without imposing their own.

Trading stories may be a more productive strategy than exchanging arguments in this particular context, because the stories about how people grew attached to their positions show how such positions often emerged as reactions to situations that were difficult to manage. The realization that the speaker was facing a challenge when their view was formed encourages interpreters to be empathetic and facilitates mutual understanding. Maybe speaker and interpreter encountered similar challenges, but they developed radically different strategies to overcome such challenges. Thus, the exchange may reveal an unexpected common ground: there was a problem to be faced. This shows how the sharing of personally significant experiences prompted by genuine curiosity may unite interpreters and speakers whereas arguments divide them. In this context, curiosity has epistemic benefits, because important information comes to the surface that helps people reach a better understanding of each other's positions. But curiosity has also moral benefits. As in the practitioner and patient interaction during a clinical encounter, also in the case of the interaction between two people with radically different worldviews, the newly acquired information about the speaker prevents the interpreter from objectifying the speaker. The speaker is not just the supporter of a bad argument, a conspiracy theorist, a climate change denialist, or an anti-vaxxer. They are people who have a variety of interests and personally significant experiences. Those interests and experiences shaped how they now think and feel about some important issues that affect everyone in society.

In the interaction between people with different identity beliefs, newly acquired information about the speaker can also the interpreter from pathologizing the speaker's mental life. The opponent is no longer deemed "irrational", "delusional", or "out of their minds", because now it is clear that the views they endorse were endorsed for a reason. For the interpreter, the speaker's views may still be wrong, misguided, even dangerous. But now they have become more understandable. The interpreter may not share the speaker's reasons or may not judge them as good reasons to hold that particular view, but has now the resources available to empathize with the speaker. As one way to attain information that can be used to enhance understanding and avoid polarization, curiosity may lead to more respectful, productive, and empathetic exchanges.

5. Conclusions

In this paper, we have endeavored to vindicate curiosity as a virtue in two ways. We have agreed with the philosophers who describe curiosity as a virtue. We have shown that, even when the disposition to seek new information fails to be an epistemic virtue, it may not amount to a vice without the intervention of another objectionable trait such

as malevolence. We have also argued that curiosity can be both an epistemic and moral virtue and that this becomes apparent when the disposition to attain new information concerns information about our fellow agents. To support the idea that curiosity has moral as well as epistemic benefits, we have presented two cases of social interactions where curiosity plays an important role: it promotes mutual understanding. In both cases, not only does curiosity benefit agents epistemically as they attain worthwhile information that they would not have attained otherwise, but it also sustains a better engagement with other agents, independent of how different they may be in status, experience, and perspective. Curiosity does not eliminate disagreement but helps us see other people as agents with a valuable perspective to share. Thus, developing and cultivating curiosity are promising ways to create the foundations for more effective and epistemically just communication.

Author Contributions: Conceptualization, L.B. and K.M.-H.; methodology, L.B. and K.M.-H.; investigation, L.B. and K.M.-H.; resources, L.B. and K.M.-H.; writing—original draft preparation, L.B. and K.M.-H.; writing—review and editing, L.B. and K.M.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were used in this article.

Acknowledgments: The authors acknowledge the editor and two anonymous reviewers for constructive feedback. They also thank the University of Birmingham Women in Philosophy group and in particular Jessica Sutherland, Ema Sullivan-Bissett, Lucienne Spencer, Ellie Harris, Fer Zambra, Valeria Motta, and Rosa Ritunnano for comments on an earlier version of the paper. Kathleen Murphy-Hollies is grateful to Christian Miller and Pilar Lopez-Cantero for insightful conversations about the topic of the paper and about virtues more generally. These discussions took place at the Honesty Project's Summer Seminar series, funded by the John Templeton Foundation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vogl, E.; Pekrun, R.; Murayama, K.; Loderer, K. Surprised—Curious—Confused: Epistemic emotions and knowledge exploration. *Emotion* **2020**, *20*, 625–641. [CrossRef] [PubMed]
2. Berlyne, D.E. A theory of human curiosity. *Br. J. Psychol. Gen. Sect.* **1954**, *45*, 180–191. [CrossRef] [PubMed]
3. Zagzebski, L. *Virtues of the Mind*; Cambridge University Press: Cambridge, UK, 1996.
4. Ross, L. The virtue of curiosity. *Episteme* **2020**, *17*, 105–120. [CrossRef]
5. Miščević, N. *Curiosity as an Epistemic Virtue*; Springer: Berlin/Heidelberg, Germany, 2020.
6. Watson, L. Curiosity and inquisitiveness. In *The Routledge Handbook of Virtue Epistemology*; Battaly, H., Ed.; Routledge: London, UK, 2018; pp. 155–166.
7. Baumgarten, E. Curiosity as a moral virtue. *Int. J. Appl. Philos.* **2001**, *15*, 169–184. [CrossRef]
8. Bergen, C.; Bortolotti, L.; Tallent, K.; Broome, M.; Larkin, M.; Temple, R.; Fadashe, C.; Lee, C.; Lim, M.C.; McCabe, R. Communication in youth mental health clinical encounters: Introducing the agential stance. *Theory Psychol.* **2022**, *32*, 667–690. [CrossRef] [PubMed]
9. Phillips, R. Curiosity: Care, Virtue and Pleasure in Uncovering the New. *Theory Cult. Soc.* **2015**, *32*, 149–161. [CrossRef]
10. Alfano, M. The Most Agreeable of All Vices: Nietzsche as Virtue Epistemologist. *Br. J. Hist. Philos.* **2013**, *21*, 767–790. [CrossRef]
11. Cochoy, F. *On Curiosity: The Art of Market Seduction*; Lira, J.T., Translator; Mattering Press: Manchester, UK, 2016.
12. *Holy Bible*, New International Version, NIV 2011. BibleGateway.com. Available online: http://www.biblegateway.com/versions/New-International-Version-NIV-Bible/#booklist (accessed on 19 July 2023).
13. Harrison, P. Curiosity, Forbidden Knowledge, and the Reformation of Natural Philosophy in Early Modern England. *Isis* **2001**, *92*, 268.
14. *Augustine Confessions*; Chadwick, H., Translator; Oxford University Press: Oxford, UK, 1991; pp. 397–400.
15. Aquinas, T. *Summa Theologiae (The Summa Theologica)*; Fathers of the English Dominican Province, Translator; Benziger Bros Edition: Cincinnati, OH, USA, 1981; pp. 1266–1273.
16. Buzzell, A.; Rini, R. Doing your own research and other impossible acts of epistemic superheroism. *Philos. Psychol.* **2022**, *36*, 906–930. [CrossRef]
17. Bacon, F. *The Advancement of Learning*; Dent, G.W.K., Ed.; Cambridge University Press: London, UK, 1962; p. 1605.
18. Hobbes, T. *Leviathan*; Gaskin, J.C.A., Ed.; Oxford University Press: Oxford, UK, 2008; p. 1651.

19. Baehr, J. The structure of intellectual vices. In *Vice Epistemology*; Kidd, I.J., Battaly, H., Cassam, Q., Eds.; Routledge: Informa, London, UK, 2020.
20. Whitcomb, D. Curiosity was framed. *Philos. Phenomenol. Res.* **2010**, *81*, 664–687. [CrossRef]
21. McCabe, R.; Skelton, J.; Heath, C.; Burns, T.; Priebe, S. Engagement of patients with psychosis in the consultation: Conversation analytic study. *BMJ* **2002**, *325*, 1148. [CrossRef] [PubMed]
22. Zangrilli, A.; Ducci, G.; Bandinelli, P.L.; Dooley, J.; McCabe, R.; Priebe, S. How do psychiatrists address delusions in first meetings in acute care? A qualitative study. *BMC Psychiatry* **2014**, *14*, 178. [CrossRef]
23. Bortolotti, L. *Why Delusions Matter*; Bloomsbury: London, UK, 2023.
24. Murphy-Hollies, K. Self-Regulation and Political Confabulation. In *Values and Virtues for a Challenging World*; Jefferson, A., Palermos, O., Paris, P., Webber, J., Eds.; Royal Institute of Philosophy Supplement, Issue 92; Cambridge University Press: Cambridge, UK, 2022.
25. McRaney, D. *How Minds Change*; Bloomsbury: London, UK, 2022.

Appendix

Paper 2: What is left of irrationality?

Murphy-Hollies, K. and Caporuscio, C., (2023), 'What is left of irrationality?', in Philosophical Psychology, 36(4), pp.808-818. DOI: https://doi.org/10.1080/09515089.2023.2186220

Philosophical Psychology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/cphp20

What is left of irrationality?

Kathleen Murphy-Hollies & Chiara Caporuscio

To cite this article: Kathleen Murphy-Hollies & Chiara Caporuscio (2023) What is left of irrationality?, Philosophical Psychology, 36:4, 808-818, DOI: 10.1080/09515089.2023.2186220

To link to this article: https://doi.org/10.1080/09515089.2023.2186220

Published online: 21 Mar 2023.

PHILOSOPHICAL PSYCHOLOGY
2023, VOL. 36, NO. 4, 808–818
https://doi.org/10.1080/09515089.2023.2186220

Routledge
Taylor & Francis Group

What is left of irrationality?

Kathleen Murphy-Hollies[a] and Chiara Caporuscio[b]

[a]Department of Philosophy, University of Birmingham, Birmingham, UK; [b]Facultät für Humanwissenschaften, Otto-von-Guericke Universität, Magdeburg, Germany

ABSTRACT

In his recent book *Bad Beliefs and Why They Happen to Good People*, Neil Levy argues that conspiracy theories result from the same rational processes that underlie epistemic success. While we think many of Levy's points are valuable, like his criticism of the myth of individual cognition and his emphasis on the importance of one's social epistemic environment, we believe that his account overlooks some important aspects. We argue that social deference is an active process, and as such can be helped or hindered by epistemic virtues and vices. With this in mind, holders of bad beliefs acquire more responsibility than is considered by Levy.

ARTICLE HISTORY
Received 6 December 2022
Accepted 22 February 2023

KEYWORDS
Rationality; social deference; identity; epistemic virtues; conspiracy

Introduction

Conspiracy theories and misinformation are widespread phenomena. Claims on which the scientific community has long reached a consensus, like anthropogenic climate change, evolution, or the efficacy of vaccinations, are disputed by unreliable sources whose alternative stories are believed by large parts of the population despite the lack of evidential support. The story according to which vaccinations are an elaborate masterplan by pharmaceutical companies to implant chips into members of the population has been repeatedly rejected by experts and yet, it remains for many a more attractive theory of the purpose of vaccinations than the mainstream view. Why would a rational agent believe such a far-fetched, convoluted and discredited story rather than one that is linear, simple and largely supported by evidence?

A popular answer to this question is that people often do not act as rational agents. Since the Enlightenment, rationality has been characterized as a largely individual process: our capacity to collect first-order evidence, weigh it appropriately and come to the most plausible conclusion, with the help of epistemic virtues such as being open-minded, tenaciously logical,

CONTACT Kathleen Murphy-Hollies KLM276@student.bham.ac.uk Department of Philosophy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK; Chiara Caporuscio caporusciochiara1@gmail.com Facultät für Humanwissenschaften, Otto-von-Guericke Universität, Magdeburg, Germany

conscientious, humble, and so on. According to this conception, accepting conspiracy theories and other bizarre beliefs is a failure of rationality that is to be blamed on individual biases, epistemic vices and reasoning errors which affect people's capacities to deal with first-order evidence. Bad beliefs are the result of irrational processes.

In his book "Bad Beliefs: Why They Happen to Good People", Neil Levy (2021) presents an alternative account that sees these bizarre beliefs as resulting from the same rational processes that underlie epistemic success. Levy argues that much of human cognition does not rely on first-order evidence as much as it does on social deference and higher-order evidence. Most people do not have the resources or background knowledge necessary to assess the first-order evidence for or against anthropogenic climate change, for example. Both the scientifically-minded individual and the conspiracy theorist need to trust second-order sources of information which have already analyzed and interpreted first-order evidence. The only difference between the two individuals is where they place this trust.

According to Levy, this means that conspiracy theories are not to be blamed on individual epistemic agents, but on the epistemic environment they are immersed in. It is a rational choice to defer to trusted sources of information on topics where the first-order evidence is too complicated to deal with yourself. However, the polluted epistemic environment we live in makes it so that the higher-order sources that are more present and vocal in the lives of many are not the ones that should be trusted. Thus, if we want to combat the rise of conspiracy theories, we need to clean up the epistemic environment first; by making it so that people who are not experts cannot exhibit expert status, and by increasing credibility signals of the scientifically supported opinion. This way, we can make it easier for people to know and recognize which higher-order sources to trust.

A lot of Levy's points are timely and well-taken. We appreciate Levy's project of showing that the average person is not a stupid irrational being, but someone who makes choices which make sense to them. We also appreciate the emphasis on just how significant and important one's social epistemic environment is, and that knowledge production is very much a fundamentally shared enterprise. However, we believe that his account can overlook some important parts of the story: the social aspect of epistemic virtues and vices, and the role of active choice in belief formation. When considering these aspects, we think that falling for conspiracy theories and bad beliefs acquires more epistemic responsibility than Levy allows.

In part one, we take a closer look at some of the examples discussed by Levy and consider how they affect what rationality, and opposingly, irrationality, mean. These examples look rational with hindsight but don't involve comprehensive understanding. Sometimes, we want more than this; we want to innovate the processes and conclusions we acquire socially

by altering them and improving them and this takes closer engagement. This is especially true when it comes to high-stakes beliefs about climate change, or the safety of vaccines. We also consider that epistemic virtues can play a more valuable role here than Levy allows. In part two, we argue that belief formation is an active process of picking sides based on one's self-conceptions, rather than a passive process where beliefs that are prevalent in our environment "happen" to us. This puts responsibility back into the picture and suggests that cleaning up the epistemic environment won't be enough to solve the problem of bad beliefs.

Part 1: Rationality as luck

Levy provides a convincing and comprehensive account of how some of our strangest beliefs and practices can in fact be understood as rational. However, we worry that in managing to rationalize such practices, we start to lose out on a useful picture of *irrationality*. We'll look at two of Levy's examples to demonstrate this. Firstly, Levy discusses the Naskapi hunters who heat a caribou shoulder over coals until it cracks, and then decide where to hunt on the basis of how the pieces fall. Secondly, Indigenous American peoples who cooked corn with wood ash or ground sea shells or lime, and so subsequently the corn did not give them Pellagra, which was a disease affecting corn-eaters elsewhere. Both practices are just "the done thing", with both populations having little grasp of the mechanisms by which they work. In the former case, the random nature of how the bone fragments fall ensures that the hunters don't fall prey to the tendency to get superstitious and see illusory patterns in which hunting spots are best. In the latter case, the added ingredients to the corn were alkalis which released the niacin in the corn, which in turn meant that the corn did not give people the Pellagra disease (caused by niacin deficiency).

Levy describes both of these practices as perfectly rational and reflective of the crucial role of culture in knowledge production. He emphasizes the severe costs if individuals break away from doing "the done thing" here and question the practices; they are less successful and risk illness or even death. Because the individual who breaks away from these practices and questions them risks so much and neglects such valuable social knowledge, it is a better epistemic position to be in to just go along with the practices even if the mechanism isn't understood. Levy describes these practices as therefore manifesting "intelligence" (2021, 49). It is a special skill of humans to imitate every step of a routine shown to them even if some steps are clearly functionally redundant. Chimpanzees will not do this, skipping out the unnecessary steps. This is, in Levy's view, to their loss because it gets in the way of accumulating very valuable cultural knowledge over time and generations, which would be

impossible for individuals working alone. However, he also says that "we owe our success to the fact that we are in some ways less – or at any rate less directly – rational animals than chimps." (2021, 45) This hints at the possibility of what is rational and what is "successful" or "intelligent" coming apart, but it is by these same processes that Levy goes on, throughout the book, to defend bad beliefs as rational.

Our worry is that these practices manifest intelligence from an external point of view, of mother nature, perhaps. They work, in the long run. But this doesn't tell us very much about people and rationality, with the latter turning into a matter of luck. Specifically, luck with regards to whether you are an Indigenous American with the custom of cooking corn with ash or shells or lime, or if you are based elsewhere and do not do this. Irrationality becomes no fault at all, but just "wrong place wrong time". At first this fits Levy's picture to some extent – people are not irrational, just their environments are unideal and either they have good customs or they don't – but Levy also allows that within the same culture, some cultural practices will be "shallow" and require straightforward imitation, whereas others will be "deep" and require *innovation* in order to achieve the valuable cultural knowledge accumulated over time.

Returning to the shells example, we want the Indigenous Americans not to question their corn-cooking practices, but for the Europeans to do so. How are we ever to know which position we are in? Levy describes the intelligence of the caribou-shoulder burning as the overriding of the human disposition to lose signal in noise by seeing illusory patterns, but the practice around corn-cooking would have *initially* been a pattern which could have been just as illusory as the superstitions which damage hunting prospects – because there is no understanding of the underlying mechanism to guide this decision-making process. These cooking practices clearly turn out to be worthwhile, but from the point of view of the human beings involved, it's a poorly understood ritual that they are "falling prey to" in the same way that the hunters would be "falling prey to" biases of superstition.

Medical professionals were in this position when investigating why rates of Pellagra were so high outside of the Indigenous American population, and asking questions about the mechanism (or, innovating) is what brought answers. This is where we want good old fashioned individual rationality to come in; in ascertaining when to question and innovate, and when to go with the flow and imitate faithfully. Investigating the underlying mechanism and ascertaining how "illusory" the pattern really is will be crucial to this. Levy acknowledges that we do sometimes respond to hints in deciphering this; if the person we are imitating seems to be acting very intentionally we are less likely to innovate. If the person we are imitating seems to be distracted or getting around another problem (their hands are full, for example), we are more likely to innovate and not straightforwardly imitate.

But this is difficult to apply to helping us know when we are dealing with shallow or deep cultural knowledge. It is difficult to apply to cases of individuals faced with the question of whether climate change is real, or whether they should support figures like Donald Trump. Levy describes how in the latter case of Never Trumpers especially, it is the social outsourcing of beliefs and falling in line with what those around us believe because they are perceived to be "people like us" (and perhaps also as prestigious) which explains how the seemingly irrational change of opinion is ecologically rational. In other words, it's a mechanism which usually gets us good and well-supported beliefs, but in these kooky epistemic environments they get us bad beliefs. We accept that these individuals often have good reasons for rejecting the mainstream view given their epistemic environment and peer group, but nevertheless the option of innovation surely doesn't go away. Levy says that innovation is appropriate for *shallow* cultural knowledge, as opposed to deep cultural knowledge, but this merely pushes the question one step along to – how do we know whether an issue pertains to shallow or deep cultural knowledge? Given that Levy appears to allow for the issue of whether to support Trump, and other beliefs which *feel* personally deep, to in fact be shallow (2021, 65) because they are abandoned relatively quickly in response to social pressure, this is a sticky problem. But importantly it surely makes it possible that beliefs regarding the truth of climate change can also count as shallow and the option of innovation remains. We are suggesting that some epistemic responsibility may come into the picture in ascertaining this; whether we should innovate or imitate, even if perfectly rational processes can lead us astray once we pluck for one of these.

The next question is likely to be, what does "innovation" look like in the face of considering whether to accept the truth of climate change? Levy warns against individualistic solutions; "doing your own research" or having epistemic virtues. We look at each of these in turn.

We accept that innovation does not have to be an individualistic affair, and can instead be just as socially embedded as Levy's picture of rational processes is. This is in contrast with his description of "doing your own research" as being very individualistic. He describes agents as facing a choice of either shrugging their shoulders, or "doing their own research" and digging into argumentation, when they come across surprising or bizarre conclusions such as that climate change is not real. He suggests that they ought to shrug their shoulders and move on (2021, 94), given the risks which the individual incurs when engaging in questioning (such as in the cases of Indigenous Americans who would question cooking corn in their traditional way, and Naskapi hunters who would question why they use caribou shoulder fragments to pick where to hunt). However, we think this is too simplistic. Firstly, both these options are individualistic; shrugging

shoulders or engaging and trying to tackle spurious arguments ourselves. But Levy criticizes only the second for being individualistic. We also do not think this option has to be individualistic. It only looks this way when we needlessly limit the time span we are looking at, to the immediate aftermath of coming across a strange conclusion and/or argument. In reality, we think there is a path between the two options of shoulder shrugging and individual research. This is something like, holding the strange idea in the back of our minds, and seeing what happens in the near future. Do you notice other people mention it? Do other people ask you about it? Do you come across specific people you think could give really valuable insight? Does it pop up on twitter or in meetings? Perhaps you follow up on that when you might not have before. We hope it's clear to see the role that other people play here in tackling a surprising new idea. But it is not an attempt to, independently, master complex expert literature or "science the shit" out of something, nor is it a passive shoulder shrug given that we already know what we and our peers think about some issue.

Zooming out, our picture is one of agents, over time, sometimes finding themselves alone with a new idea and perusing argumentation behind it, reflecting on how it strikes them, and being in moments where they have no choice but to be individualistic. But at others, they draw on the thoughts and ideas of others consciously or subconsciously, perhaps to then think about privately again later. In this cycle, there are individualistic moments which Levy captures but eschews, whereas we think they can still have a part to play in a broader process which draws on social influences at other times.

We think something similar is the case with Levy's account of epistemic virtues. He thinks that epistemic virtues are not as risky as doing your own research, but still worries that they are too individualistic to help - "they appear to aim to bring us each to inculcate the virtues in ourselves and then, guided by our intellectual excellences, to tackle hard problems largely on our own" (2021, 91). He would prefer something which better enables apt deference to others. However, we see a much closer link between epistemic virtues and exactly this – apt deference to others. Things like open-mindedness, humility, arrogance, sociability, can all have a significant role to play in ensuring apt deference to others. If we are arrogant, we are unlikely to take much of what anyone else says seriously. If we are humble, we are more likely to take seriously what we hear from others, and not just people who look like us or are familiar to us. If we are sociable, we're likely to be in more, and more intimate contact with a wider variety of other people from all walks of life, and therefore be more likely to come across lots of valuable information from them. In many ways, the virtue epistemologist has to battle the same problems as the virtue ethicist in accepting that unideal environments – particularly unideal social environments – do place individuals at significant disadvantages in their epistemic, or ethical,

development. They both hope that agents will respectfully defer to others, have relevant trustworthy epistemic institutions – or moral role models – available to them, and have relevant educational experiences to learn from. They are both concerned not just with ways of analyzing and interpreting first-order evidence (of what to believe or how exactly to act), but of being well disposed such that you're in a good position to defer aptly to others when needed and bring what is learnt there to bear on relevant situations.

So, we do not see "doing your own research" or virtue epistemology as individually as Levy seems to. Although this lets into the picture all the ways that unideal social environments can create bad beliefs, much of Levy's description of which we are on board with, we still also think there are some opportunities here for slightly more independent choice and rationality to be exercised.

Part 2: Beliefs, action and responsibility

Levy's take on rationality hinges on some underlying assumptions about how we form and sustain beliefs. Beliefs "happen" to people; the title itself implies that belief formation is not an active process of choosing the most rational option after analyzing the available first-order evidence, but it is a somewhat passive process that can be reliably predicted given certain environmental factors. If the prevalent sources of information that hold epistemic authority in my environment says that x, the belief that x will likely "happen" to me. Rationality does not need to involve active thinking or choosing, but imitating practices and deferring to one's social environment. If the epistemic environment is polluted and filled with misinformation, the otherwise rational act of social deference will fail and conspiracy theories will proliferate.

This has important consequences for Levy's proposed solution to the proliferation of conspiracy theories and bad beliefs. If the environment is the primary force determining beliefs, we can artificially generate better beliefs by cleaning up the epistemic environment. Specifically, by making clear which are the mainstream, scientifically supported views and nudging people towards credible positions, we will counterbalance the rise of conspiracy theorists and produce more successful epistemic agents.

Levy does capture something important here. The environment and the available evidence are certainly reasons for epistemic success or failure. Someone growing up in a family of scientists is probably less likely to succumb to conspiracy theories than someone whose social bubble is made entirely of flat-earthers. Having credible information available and easily recognizable is an important prerequisite to forming good beliefs, and our current epistemic environment is not ideal in this sense. Unless one has learned a fairly complicated set of skills to help them recognize which

sources to trust, the amount of contradicting and (at least at first sight) credible-looking information around is not easy to navigate.

However, talking about beliefs in this way ignores another very important aspect of belief formation. We do not only form beliefs by passively absorbing information – we actively select and choose what or who we want to trust in based on our personality, epistemic strategies, identity and core beliefs. Making it obvious what views are mainstream and scientifically credited will only help people who have already made the choice of trusting mainstream, scientific views.

A devoted Christian will likely keep believing in creationism even if they end up in an epistemic environment where the most prevalent sources of information say otherwise, like a largely non-religious city or a science class. A left-leaning scientist will keep believing in climate change if they move to a very conservative town where everyone thinks green politics are a hoax. Both epistemic agents are perfectly capable of ignoring the mainstream position in their current environment in favor of a minority position. More importantly, they do so in spite of all pointers of epistemic authority (it being taught at school, for example) because of a background choice about their values and who they are. They know who the epistemic authority is in their social environment; they just decide to reject it because they perceive it as conflicting with or not speaking to their own values. In extreme cases, they might come to reject any view that comes to their attention that is labeled as epistemically authoritative, even if they know nothing about it, precisely because it is presented as the consensus in an epistemic environment they feel strongly averse to. They think that 'this is the narrative of the system; I do not fit in the system; so this is not my narrative'.

This is in contrast to Levy's description of even deeply held (or so it seems) beliefs about the self being shallow and often being outsourced – as he suggests is the case with the "Never Trumpers" who denounce Trump but then ended up supporting him, in line with the rest of the social group they saw themselves as being a part of. Instead of having robust inner models of ourselves and our beliefs, we off-load these onto the outer world and tend to more so just respond to triggers when the time comes. In the most stark example of this, choice blindness studies show individuals explaining their recently expressed belief or choice x when prompted to by researchers, even though they actually expressed belief or choice y. So, they used this prompt by researchers as a trigger which told them what they actually thought/believed, which demonstrates how shallow and impoverished our inner models are.

However, an alternative interpretation of this is that we actually see quite how tightly people cling to inner models of themselves. Individuals react to these prompts in this way because of some of their other self-related beliefs.

Particularly, for example, that they are consistent, that they know what they said a minute ago, that they are the ultimate authority on their own views. This is a strange scenario in which deeply held beliefs about one's own consistency does cause inconsistency, given that evidence to the contrary is being overlooked. But this is still in the light of enjoying a particular self-image.

So, for similar reasons, a Flat Earther will likely not abandon their convictions just because the epistemic environment gets depolluted and the mainstream view is clearly indicated. This is because this depolluting doesn't contribute anything to the self-relating beliefs which the agent practically enjoys having, because they speak to their being consistent but also interesting and contrasting. Even more than religion or science, many conspiracy theories thrive in an "us versus them" dynamic. Such polarization cannot only be explained by looking at a polluted epistemic environment and the difficulty in telling the scientific information apart from the unfounded sources. It is, in the first place, an active choice to pick one side rather than the other. We will briefly talk about which factors might influence this choice, and the consequences for moral and epistemic responsibility.

We believe that this is an area where epistemic virtues can actually be helpful. In particular, for example, it seems that a common epistemic vice among conspiracy theorists is a need for uniqueness (Douglas et al 2019, 9) – a desire to place oneself above others in terms of one's epistemic capacities and knowledge, to hold the essentially contrasting minority position and defend controversial views in order to appear different from the majority. This could be thought of as a sort of, propensity to be an "epistemic special snowflake" and emphasizes how something that looks quite a lot like a vicious trait, has knock-on effects for aptly deferring to good sources and paying them the attention they deserve.

Accepting some wacky beliefs that might contradict one another at a superficial level is a worthy sacrifice to maintain stability in one's core beliefs at a deeper inner level. People who feel rejected and alienated by the system are much more likely to develop conspiratory beliefs (Pierre, 2020). Losing faith in the system develops into actively researching narratives outside of it, which further fosters the feeling of opposition: we are smarter than them, we know something that they are trying to hide. This opposition is not only epistemic, but is often also experienced as moral, social and political. One recent example are COVID deniers (Bisiada, 2021). Refusing to comply with governmental policies led to them being pointed to as responsible for the continuation of the pandemic by people who were practising social distancing, wearing masks and getting vaccinated. In turn, being painted as ignorant led to further grudge and distrust toward the official narrative from people who were skeptical, uninformed or just

unable to deal with the social and economical consequences of social distancing.

The deeper issue with many instances of adherence to conspiracy theories is not to be found in misinformation itself, but in the reasons why misinformation is picked and believed. In all these cases, mainstream views are rejected not because it's hard to know which view is the epistemically authoritative view in an epistemically polluted environment, but precisely because of the signals of authority they display. If this is true, cleaning up the epistemic environment will not have the impact that Levy hopes for – highlighting with banners and pointers what the consensus of epistemically authoritative sources is will not make these people gain trust in those authorities, rather it will make it easier to recognize which positions to reject, because adopting them does not emphasize and enhance the individual's self-conceptions. Without deeper work on the socio-political environment to stop fueling "us versus them" narratives, increase trust in science and prevent groups from feeling marginalized, these efforts to contrast the rise of bad beliefs will be limited.

Conclusion

Where does this leave rationality and responsibility? We have argued that social deference can be rational not in virtue of a passive mechanism of absorbing or repeating the prevalent practice or opinion within one's immediate environment, but in virtue of a powerful active component: one's capacity to adequately pick who to trust. This choice is strongly influenced by one's identity and can be led astray by epistemic vices like a need for uniqueness, and alternatively helped by epistemic virtues like humility, and therefore it is not free from epistemic responsibility. At the same time, Levy is right in arguing that to change people's beliefs, we need to start from changing their environment. Rather than a simple epistemic depolluting, though, we suggest that social changes are necessary in the first place to prevent people from developing anti-scientific identities and consequently picking anti-scientific beliefs.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Bisiada, M. (2021). Discursive structures and power relations in Covid-19 knowledge production. *Humanities and Social Sciences Communications*, *8*(1), 1–10. https://doi.org/10.1057/s41599-021-00935-2

Douglas, K. M., Uscinski, J. E., Sutton, R. M., Cichocka, A., Nefes, T., Ang, C. S., & Deravi, F. (2019). Understanding conspiracy theories. *political Psychology*, *40*, 3–35. https://doi.org/10.1111/pops.12568

Levy, N. (2021). *Bad beliefs: Why they happen to good people*. Oxford University Press.

Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, *8*(2), 617–641. https://doi.org/10.5964/jspp.v8i2.1362