Topoi (2020) 39:229–241
https://doi.org/10.1007/s11245-018-9611-z

Gaslighting, Confabulation, and Epistemic Innocence

Andrew D. Spear[1]

Published online: 8 November 2018

Abstract
Recent literature on epistemic innocence develops the idea that a defective cognitive process may nevertheless merit special consideration insofar as it confers an epistemic benefit that would not otherwise be available. For example, confabulation may be epistemically innocent when it makes a subject more likely to form future true beliefs or helps her maintain a coherent self-concept. I consider the role of confabulation in typical cases of interpersonal gaslighting, and argue that confabulation will not be epistemically innocent in such cases even if it does preserve a coherent self-concept or belief-set for the subject. Analyzing the role of confabulation in gaslighting illustrates its role in on-going interpersonal relationships, and augments already growing evidence that confabulation may be quite widespread. The role of confabulation in gaslighting shows that whether confabulation confers epistemic benefits (and so is epistemically innocent) will depend greatly on the interpersonal context in which it is deployed, while whether a coherent self-concept is epistemically beneficial will depend to a great extent on the content of that self-concept. This shows that the notion of an epistemically harmful or beneficial feature of a cognitive process can and should be further clarified, and that doing so has both theoretical and practical advantages in understanding epistemic innocence itself.

Keywords Confabulation · Gaslighitng · Epistemic innocence · Peer-disagreement · Epistemic benefit

This essay is about the role of *confabulation* in prototypical cases of *gaslighting*, and about whether such confabulation is *epistemically innocent*. In Sect. 1, I introduce Kate Abramson's analysis of gaslighting, and argue that gaslighting has an essential epistemic dimension. I then argue (Sects. 2 and 3) that confabulation plays a central role in many prototypical cases of gaslighting. The case of gaslighting shows how confabulation can play a role in ongoing interpersonal relationships. I take this to be interesting in its own right, but also to suggest that confabulation may be even more pervasive than is already thought. Finally, in Sect. 4, I consider whether or not the confabulation that occurs in prototypical gaslighting is epistemically innocent as understood in recent work by Lisa Bortolotti and others. I argue that confabulation in the context of gaslighting confers positive epistemic harms rather than benefits, and so is not epistemically innocent. This is not a challenge to the notion of epistemic innocence. Rather, the reasons why confabulation in the context of gaslighting fails to be epistemically innocent themselves point to specific ways that the component notion of an *epistemically beneficial* cognitive process can be further clarified, which is important for both theoretical and practical reasons.

1 Gaslighting

The term 'gaslighting' originates with Patrick Hamilton's 1938 play, "Gaslight", and has more recently become a fixture in cultural discourse,[1] featuring in publications in self-help psychology (Stern 2008) and political commentary (Carpenter 2018). The idea has been treated in academic publications first in psychoanalysis (Calef and Winshel 1981), and more recently in philosophy by Kate Abramson (2014). I will follow Abramson's analysis closely here. However, Abramson focuses on the role of manipulation in gaslighting, and downplays the role of epistemic factors such as testimony, evidence, and reasons. I think these latter are equally essential to a complete understanding of the

✉ Andrew D. Spear
speara@gvsu.edu

[1] Grand Valley State University, 1 Campus Drive, Allendale, MI 49401, USA

[1] A casual Google search for 'gaslighting' turns up 3,470,000 hits, more even than confabulation, which comes in at 1,130,000.

phenomenon. In this section I introduce my preferred characterization of gaslighting, discuss two central examples of gaslighting, and explain my difference with Abramson concerning the epistemic dimension of gaslighting.

1.1 Characterizing Gaslighting

Gaslighting involves (i) the attempt by the gaslighter to undermine his victim's self-trust: her conception of herself as an autonomous locus of experience, thought, and judgment.[2] The gaslighter's (ii) motivation is a strong desire to neutralize his victim's ability to criticize him and to ensure her consent to his way of viewing things (specifically with regard to issues relevant to the relationship, perhaps in general), and thus to maintain control over her. The gaslighter (iii) pursues this goal by means of a strategy of manipulation, fabrication, and deception that (iv) specifically relies upon his victim's trust in him as a peer or authority in some relevant sense.[3]

The most distinctive feature of gaslighting is that it is not enough for the gaslighter simply to control his victim or have things go his way: it is essential to him that the victim herself actually come to agree with him (Abramson 2014, pp. 10–12). Gaslighting is thus distinct from (though not unrelated to) silencing, as well as from creating an environment where *everyone else* believes the victim is wrong, and also from creating a situation where the victim has no choice but to acquiesce, even while *not* agreeing or seeing things his way.

From the standpoint of social psychology, it makes sense to think of gaslighting as a type of abuse. The motives of the gaslighter are consistent with "power and control" motives typically ascribed to abusers (Wagers 2015). Like other forms of abuse, gaslighting is also typically persistent, and deployed over a long period. Also, abusers are often characterized as having difficulty identifying, managing and expressing emotions, and as suffering from a strong sense of vulnerability and low sense of self-worth (Smith 2007; Wagers 2015). Someone with a psychology of this sort is likely to be motivated to gaslight insofar as a partner who (as a result of gaslighting) assents to the gaslighter's view of things is unlikely to offer emotional or other challenges. I will consider additional similarities between gaslighting and other types of abuse in Sect. 3. However, so far as I am aware, there do not exist empirical studies focusing specifically on the phenomenon of gaslighting. This being the case, in the next section I will consider two examples of gaslighting that have played an important role in recent discussions in order to illustrate the phenomenon.

1.2 Gaslighting: Two Cases

In the 1944 film *Gas Light*, Paula (Ingrid Bergman) is the victim of gaslighting on the part of her husband, Gregory (Charles Boyer).[4] Unbeknownst to herself, Paula is in possession of priceless jewels that belonged to her aunt, a famous opera singer mysteriously murdered when Paula was a girl. Gregory, in fact the murderer of Paula's aunt and now bent on gaining possession of the jewels, romances and marries Paula, then convinces her to return to London to live in her aunt's old house. Confident the jewels are in the attic, Gregory "goes out on business" often, when in reality he goes to the attic by a back staircase. His use of the gaslights in the attic dims the gaslight in the house. Paula sees this, and often hears mysterious footsteps in the attic. Gregory repeatedly assures her that she is imagining things, suggests that she is "over-tired", and questions her memory and perceptions. He arranges things so that it appears that she has been stealing, hiding, or moving objects in the house. When confronted, Paula has no recollection of having done this, which only confirms Gregory's "suspicion" that she is mentally unstable. As a result, Paula's confidence in her own judgment and mental faculties gradually deteriorates. Gregory's goal is to drive her to the point where she believes she is mad so that he can get her out of the way in order to search for the jewels more efficiently.

A second example comes from the1952 Film *Pat and Mike*, and is used extensively by Abramson (2014). Pat is an aspiring female golfer, while her fiancée, Collier, wants her to give up her golf career so they can get married and she can assume wifely duties of household and children. After a close tournament loss in which Collier's less than supportive approach has played a clear role, Pat finds herself in crisis. Her confidence and life-goals shaky, Collier takes advantage of the opportunity to press her to give up her

[2] Here and throughout I use 'she', 'her', etc. to refer to the victim of gaslighting and 'he', 'his', etc. to refer to the perpetrator. This usage helps to avoid pronoun-confusion throughout, and fits with the situations in key examples of gaslighting (the films *Gas Light* and *Pat and Mike*) and with the feminist slant of Abramson's 2014 analysis of the phenomenon. In adopting this usage, I do not mean to take a stand concerning whether gaslighting is particularly or predominantly perpetrated by men against women. I think that the agent and patient of gaslighting can be mixed in any number of ways in terms of gender (men to women, men to men, women to men, etc.) and other social relations. Who gaslights who predominantly or the most, in terms of social category, is an empirical question that I expect depends rather heavily on social and cultural factors. The analysis being offered here should apply regardless.

[3] Gaslighting is not limited to intimate partners in a relationship. As characterized here, gaslighting is possible in many interpersonal contexts involving trust or authority, such as employee-employer relationships and relationships amongst peers of various sorts (friends, co-workers, fellow students, etc.).

[4] The film is itself based on Patrick Hamilton's (1938) play of the same name.

career (something Pat clearly does not want) and marry him ("Why don't you just let me take charge!" says Collier at one point). During their conversation, Collier refuses to express support for Pat, tacitly but persistently suggests doubt in her ability to make judgments for herself, and issues the subtle threat that her failure to see things his way might result in the end of their relationship ("Just make sure you don't think it under…" Collier says, then abruptly shuts down the conversation). Whereas Gregory's gaslighting of Paula is about gaining control over her by means of providing her with significant but false evidence against trust in her own agency, Collier's gaslighting of Pat attempts to undermine her sense of her own agency largely by marshalling her emotional commitment to him (as well as her own diminished self-confidence), but with the clear goal of ensuring that she comply with his vision of their future together.

1.3 Abramson on Gaslighting

On the basis of examples such as these,[5] Abramson argues that gaslighters are individuals who cannot tolerate even the possibility of disagreement with or criticism of their way of viewing things, at least not from certain individuals (friends, loved ones, romantic partners, etc.). The purpose of gaslighting, she suggests, is not only to neutralize particular criticisms, but to eliminate the very possibility of such criticism by undermining the victim's conception of herself as an autonomous locus of thought and judgment.[6]

On Abramson's view, gaslighting essentially involves manipulation. It involves the gaslighter establishing a context in which the victim's standpoint has been challenged as untrustworthy or dysfunctional (by responding with something such as "that's crazy", or "you're not seeing this right at all!"), then issuing a demand to the victim (that she see things the way the gaslighter sees them) where the victim's motive for assent to the demand comes in the form of implicit or explicit manipulative threats on the part of the gaslighter (2014, pp. 14–15). Thus, Collier, having already called into question Pat's grasp of the situation, issues his demand to Pat, and tacitly makes it clear that if she fails to see the situation properly, he may end the relationship.

Abramson sees the manipulation in gaslighting as central. (2014, p. 16). Rather than enlisting the victim's consent by offering her a genuine reason with normative force as motive, the gaslighter enlists her consent by offering her a threat to something she cares about, where the threat doesn't have genuine normative force. So, in the case of Pat and Collier, the reason that Pat is supposed to stop trusting herself and give up her career is simply because of the fear she feels at the possibility of their relationship ending. Yet, breaking off a relationship completely is not a *fair* response to someone who resists one's unilateral demand that they give up on something important to them: the reason Collier offers doesn't have genuine normative force. It only *seems* or feels like it does to Pat because of her emotional investment. Collier's pretending to offer reasons with genuine normative force when he is really just playing on Pat's emotional commitment and insecurities is what makes this a type of manipulation.

Abramson is critical of the idea that important aspects of gaslighting are epistemic. She rejects the idea that the gaslighter is claiming a special epistemic status (as knowing better than his victim), or that mere lying plays a defining role in gaslighting. She points out that the gaslighter himself is not a sincere interlocutor concerned with getting at the truth (2014, p. 13). Further, she argues that the gaslighter typically uses assertions and manipulative threats rather than offering reasons to his victim, and that the victim herself only has "motive to assent" to the gaslighter in virtue of the personal or professional harm that he threatens (tacitly or explicitly) to cause her (2014, p. 15).[7] While I agree the gaslighter is typically insincere in these ways, I disagree with Abramson concerning her apparent view that these and other epistemic factors are not relevant, central even, to understanding gaslighting.

In typical cases of gaslighting, the gaslighter wants the victim to accept his views about two things. First and foremost, he is attempting to convince her that she should not trust herself: that she is "crazy" or "oversensitive", that her grasp of the situation is too defective for her to make reliable judgments about it, or that her judgments themselves can't be trusted. Second, there is typically also a specific further claim that the victim is supposed to accept: that she has been moving things around in the apartment and then forgetting, or that Pat should give up her career and marry Collier. The gaslighter thus, arguably, has a twofold project. There is the ongoing project of convincing his victim that she is, in general, not competent to make certain (or most) kinds of

[5] As well as others; Abramson presents an array of similar cases meant to illuminate the phenomenon in her (2014).

[6] As Abramson puts it, "…he [the gaslighter] aims to destroy the possibility of disagreement by so radically undermining another person that she has nowhere left to stand from which to disagree, no standpoint from which her words might constitute genuine disagreement." (2014, p. 10).

[7] As Abramson puts it, "…he [the gaslighter] isn't in the first instance claiming for himself a epistemic authority (I see this rightly, you don't)…what he's doing is issuing a demand that one see things his way…this isn't a case of, for instance, testimonial credence (i.e. the gaslighter isn't asking his/her target to take it on testimony that it's true that "that's crazy"). If that were the scenario, there'd be no explanation for the gaslighter's use of manipulative threats (implicit or explicit). It's the explicit or implicit manipulative threats…that give the target anything like motive for assent" (2014, p. 15).

judgments and so should not trust herself, and then there is the appeal to the results of this project as a sort of premise to compel compliance or ward off critique from the victim on particular occasions. On Abramson's view, gaslighting is only happening if the gaslighter attempts to motivate his victim to accept these attitudes toward herself and what he wants by means of manipulation. Yet it seems clear that gaslighting at least *can* be about the epistemic status of the gaslighter and about giving the victim primarily epistemic reasons to distrust her own experience and judgment. This is precisely the scenario of Gregory and Paula in the 1944 film. It is because Paula *trusts* Gregory as a reliable judge of how things are, and because he systematically manufactures plausible evidence that her faculties are untrustworthy, that she is brought to the point where she is no longer confident in her own judgments. This doesn't preclude his use of more manipulative appeals to Paula's emotions and insecurities at various points, but even when the reasons he gives her are purely (fabricated) epistemic ones, I see no reason to say that what he is doing isn't gaslighting.

The point just made about Paula's *trust* in Gregory is crucial, for it indicates a broader respect in which *all* gaslighting involves issues of epistemic status and trust. Whether the gaslighter challenges his victim's self-trust primarily manipulatively, by playing upon her emotional commitments or insecurities, primarily epistemically (by offering reasons or relying on his perceived credibility), or by means of some mix of these two strategies (probably the typical case), from the victim's standpoint his challenges to her epistemic agency place her in a situation of epistemic controversy or epistemic peer disagreement, where what she and her gaslighter disagree about is precisely the trustworthiness or reliability of her own cognitive faculties (Christenson 2009, p. 760). He maintains that she is incapable of seeing and so responding to the situation properly, and thus that she should *trust him* rather than her own grasp of what is going on. Even if he doesn't say it in just these terms, by making the assertions that he does concerning his victim's grasp of the situation, the gaslighter is, contra Abramson, laying claim to a special or privileged epistemic status relative to his victim (even if this is not *all* that he is doing). However, from the victim's own standpoint, it seems her cognitive faculties are in good order. On the plausible assumption that it is default rational to believe that one's cognitive faculties are basically reliable (see Zagzebski 2012, Chap. 2), the victim must adjudicate the question of whether her gaslighter's behavior and say-so constitute sufficient defeating reasons for her to downgrade or abandon entirely her own epistemic self-trust (her confidence in her cognitive abilities),[8] or whether his claims instead constitute grounds for trusting *him* less and so downgrading her confidence in him.[9] Indeed, part of the vertigo experienced by victims of gaslighting seems to be explainable precisely by the fact that it is not possible to provide a non-question-begging justification for relying on one's own cognitive faculties once they have been called into question in the way that the gaslighter does (Zagzebski 2012, Chap. 2). Whether the victim capitulates or not, what is crucial is that the dilemma the gaslighter's gaslighting confronts her with has an essential epistemic dimension having to do with trust, status, and credibility. It is the challenge to the victim's epistemic self-trust itself, essential to all gaslighting, and the correlated (if tacit) assertion of a kind of epistemic privilege on the part of the gaslighter, that gives the phenomenon of gaslighting an essential epistemic dimension.[10]

In short, while I agree with the lion's share of Abramson's analysis of gaslighting, I think that she is wrong not to recognize the possibility of offering false or misleading reasons as a *possible* strategy for the gaslighter, and not to recognize issues of trust, epistemic status, and the epistemic dilemma confronted by the victim as essential dimensions of *every* case of gaslighting. I agree that manipulation plays

[8] It seems clear that there are cases where the reasonable thing for a subject to do is accept that her cognitive agency has been impaired on the basis of testimony and other evidence from a trusted friend or authority (perhaps a close family member takes one aside and suggests that one appears to be suffering from early stages of Alzheimer's, providing specific instances in recent months as partial support). In gaslighting, the gaslighter presents himself as such an authoritative source of information, and the victim must decide whether his authority or her own is more trustworthy.

[9] I do not claim that the victim goes through all of this in these *exact* words, or that her thought process is anything so deliberate or conscious as this. However, she does take her gaslighter to be basically trustworthy, and trustworthy not just in a moral, but in an epistemic sense. Emotional attachment is not the same as trusting someone as basically epistemically credible. It is perfectly possible to love or be deeply invested in someone, and yet to be confident that they are not a reliable judge on many or most matters. Even if the victim has never really distinguished these two questions, the gaslighter's challenge to her grasp of the situation *de facto* distinguishes them. She must take him to be minimally sincere and reliable in an epistemic sense in order to take his challenge to her grasp of the situation seriously. If she did not think he was epistemically trustworthy or had some reason to doubt that he was, she might still go along with what he wanted for moral or prudential reasons, but she would simply have no good reason, even from her own standpoint, to change her mind about her own credibility and agency.

[10] It is because of the way in which trust is at issue in all gaslighting that I explicitly include in my characterization above that the gaslighter (iv) relies upon his victim's trust in him as a peer or authority in some relevant sense. Even a stranger could gaslight someone, as long as the person being gaslighted believed that the stranger was essentially informed and sincere in his claims: as long as she trusted him. But without this basic threshold of trust, gaslighting simply wouldn't get off the ground. If the victim doesn't trust the person who is attempting to gaslight her, doesn't think that he is a basically sincere source of accurate information or evaluation, then the gaslighting project gets no traction.

a fundamental role in gaslighting, but this is compatible with acknowledging the epistemic dimensions just outlined. Indeed, a pre-condition for successful manipulation is precisely that the victim have *some* degree of trust in the gaslighter's credibility. It is the fact, as I take it, that gaslighting involves both a manipulative and an epistemic dimension that helps make clear the role played in it by confabulation. I introduce and discuss confabulation in the next section, and explain its role in gaslighting in Sect. 3.

2 Confabulation

Confabulation is a multi-faceted phenomena being studied from a large number of disciplinary perspectives (Hirstein 2009a, b). In psychology, confabulation first had to do with false reports about memory issued by patients suffering from neurological disorders, though confabulation is recognized to occur in a range of pathological cases including reports about emotions, perceptions, somatic awareness (particularly of paralyzed limbs), and intentions (Kopelman 2010; Hirstein 2009b). In the case of memory, patients might report, when asked what they had done during the day, a series of events or activities that are inconsistent or implausible relative to each other, or relative to the context in which the question was asked (though the patient has been in the hospital for days or weeks, she may report having gone to see friends or visiting a city a significant distance away). In the case of perceptual awareness, patients with a paralyzed limb may deny that the limb is paralyzed and even tell stories about why they "choose not to use it" when asked.[11]

While confabulation originates in the context of psychology and pathology, researchers have been quite successful in prompting confabulation in subjects with no underlying neurological lesions or damage. Nisbett and Wilson famously prompted subjects to confabulate by offering them a row of retail objects (e.g. a row of socks) to choose from, then asking them *why* they chose the object that they did (rather than the others). In one version of this study subjects were offered four identical pairs of nylon stockings to choose from, while in another they were offered four different nightgowns (Nisbett and Wilson 1977, pp. 243–234). In both versions of the experiment, subjects consistently chose the object furthest to the right (the right-most pair of socks was particularly popular), yet no subject cited the object's *location* as their reason for choosing. Instead, the subjects confabulated. When asked why they chose the object they had, they generated plausible but, when seen in light of the data generated by the experiment, false reasons for their choices (Nisbett and Wilson 1977). Responses elicited in some social psychology experiments have also been labelled confabulatory. For example, Jonathan Haidt presented subjects with a vignette describing consensual incestuous sex between two adult siblings and asked them about the moral acceptability of what the individuals described had done (Haidt 2001). A majority of subjects objected morally to the behavior described in the vignette. When asked to explain their judgment, they persisted in pointing to possible harm to future off-spring as a reason even when the case had been described in such a way as to rule this possibility out completely. These responses seem confabulatory insofar as they involve subjects sincerely asserting as reasons things that, given the evidence they themselves are aware of (the vignette) are unsupported.

Max Coltheart has recently proposed a basic framework for understanding confabulation of both the pathological and the non-pathological sort (2017). On Coltheart's view, all confabulation has the following two things in common. First, a subject is asked a question concerning something about which it is generally expected people are able to give answers. However, in cases that elicit confabulation the subject does not know the answer. This places the subject in a quandary. Despite a strong felt need for an answer to the question, no answer is forthcoming. This sets the second part of the confabulation process into motion. Coltheart argues that there is a "deep human drive for causal understanding", one that motivates subjects to attempt to make sense out of or provide an explanation for things.[12] We want to understand things, and in cases where we don't, the desire to understand sets abductive reasoning (inference to the best explanation) into motion in order to provide us with a more or less satisfying understanding (2017, pp. 67–68). On Coltheart's view, confabulation is just the ordinary drive for causal understanding deploying abductive reasoning in a situation that is abnormal. Lacking direct access to the answer, the subject considers some part of the evidence that she does have relevant access to and formulates a plausible (from her standpoint) response based on it: she confabulates.

Ema Sullivan-Bissett, in her discussion of confabulation and implicit bias, offers an account that is consistent with Coltheart's, but more nuanced. She identifies five major features of confabulation: confabulations are (1) false or ill-grounded, (2) offered in response to a question, (3) they have

[11] The most prominent neurophysiological explanation of confabulation is that it is the result of two related or interconnected pieces of cognitive damage: (i) damage to some relatively modular cognitive process (memory, perception, somatic awareness, object-recognition, emotional response) coupled with (ii) damage to some less-modular executive system that would, normally, facilitate recognition of the damage/malfunction in (i). This leaves the subject with a cognitive impairment that she is unable to recognize, leading her to confabulate or fill in gaps when questioned (Hirstein 2005, 2009b, pp. 4–5).

[12] Coltheart relies here in particular on the work of Alison Gopnik (2000).

a motivational component (subjects have desires or interests that push them to confabulate), (4) they fill in a "gap" in the subject's cognitive system, a perceived dissonance or incongruity in the subject's understanding, and (5) they are reported with no intention to deceive (Sullivan-Bissett 2015, pp. 551–553).

Sullivan-Bissett offers the example of Roger, a university professor who has an implicit bias against women: he tacitly thinks they are worse colleagues and scholars than men. Roger is also on the committee reviewing CVs for a position open in his department. Because of his implicit bias, he declines to extend an offer for campus interview to Katie, a woman whose CV is, as a matter of fact, as good as or slightly better than those of the men who were selected. When asked why he did not extend an offer to Katie, Roger responds that he believed her CV was not as good as those of the other candidates. The question at issue is *why* he believes this. If Roger says "because the CV is worse than that of other candidates", then he is confabulating. Roger is wrong in two ways here. First, he is wrong about the real *cause* of his belief that Katie's CV was worse than those of the other candidates: he says the cause is that it *is* worse, but the fact is that his implicit bias is what caused this belief. Second, he is wrong in his apparent assessment that the evidence available to him rationally supports or justifies the belief that Katie's CV is worse than those of the other candidates.

Roger here fulfills all of Sullivan-Bissett's conditions. His belief is (1) false and ill-grounded. (2) He was asked a question, and issued a response. (3) We can suppose that Roger has strong motivations not to believe that Katie's CV was as good as those of other candidates, stemming perhaps from strong desires to be (and think of himself as being) objective and unbiased, in particular not to be a sexist person. His response is thus motivated. (4) His response also fills a gap in his available evidence. Roger knows he believes Katie's CV is worse and knows he did not invite her for an interview. He also knows he has seen her CV. Were he to admit that her CV is as good as or better than those of the other candidates he would encounter cognitive dissonance, a "gap", that would force him to search for other hypotheses to explain why he holds the belief he does (among which alternative hypotheses, of course, would be a failure of objectivity on his part, or that he is implicitly sexist). Finally, (5) he responded sincerely. So far as Roger is able to tell, his response is truthful and accurate.

I think that Sullivan-Bissett's discussion gets both confabulation and Roger essentially correct. I would emphasize, as a further elucidation of confabulation, that both pathological and non-pathological confabulation do seem specifically to involve "explanatory inference" or inference to the best explanation. This is the point made by Coltheart above, and Sullivan-Bissett's discussion of false and ill-grounded beliefs in confabulation specifically mentions "ill-grounded explanations", suggesting that she too sees explanatory abductive inference as central to confabulation (2015, p. 551). I think that the "gap" that subjects who confabulate confront is typically a form of dissonance resulting either from incoherence within their belief set or from being confronted with new evidence that challenges something they already believe. In the philosophy of science, it is commonly recognized that disconfirming evidence forces adjustment's in a subject's belief-structure, but rarely strictly requires the rejection of any one specific belief since auxiliary hypotheses can always be rejected (or formulated, depending on the case) instead.[13] Even the question of which of two competing hypotheses is "the most" rational in light of available evidence may not always have a clear answer (Kuhn 1977).

I do not think that hypothesis testing in science and hypothesis testing in ordinary life are that different (ordinary hypothesis testing is, if anything, typically less rigorous), and it is thus not difficult to see how a subject who is *motivated* to see (and not to see) things in a certain way might confabulate, that is, fill in a gap in his belief structure with a hypothesis that is more psychologically desirable rather than with the one that a cool objective assessment of the evidence might require.[14]

In what follows I will presuppose Sullivan-Bissett's analysis, adding the specification to her condition (1) that confabulatory beliefs are typically ill-grounded in the sense of being an abductive inference that ignores or misrepresents important features of the subject's total situation, understanding, or evidence. The next section examines how confabulation is involved in prototypical cases of gaslighting.

[13] The point is sufficiently common as to be included in introductory texts on the history and philosophy of science, for example DeWitt (2010, Chap. 4.) Thus, neo-Aristotelian Ptolemaists took the fact that an object dropped from the top of a tower falls in a line parallel to the tower and strikes the ground directly beneath where it was dropped as decisive evidence that the Earth did not rotate on its axis, while Galileo took this apparently disconfirming reasoning to be both question-begging, and also explainable in terms of the Copernican hypothesis of a moving Earth (Galileo 1632/1989, pp. 61–81).

[14] This point is further driven home by so-called "theory-theory" views of self-knowledge according to which subjects typically come to know their own beliefs, desires, and commitments not by some type of direct introspection, but rather in the same way as they do those of others: by formulating hypotheses based on observed speech and behavior (Carruthers 1996; McKinnon 2003). If self-knowledge too is a largely abductive process subject to underdetermination and to being compromised by motivated cognitions of various sorts, then the scope of confabulation, both pathological and non-pathological, in human cognition may be very great indeed (Carruthers 2010).

3 The Role of Confabulation in Gaslighting

In Sect. 1 I proposed an analysis of gaslighting according to which the phenomenon has an essential epistemic dimension. The victim of gaslighting finds herself confronted with a dilemma concerning who to trust or view as more credible: herself or her gaslighter. This is not merely a prudential or normative question, but also an epistemic one. It should be clear that there is also a very strong motivational component in gaslighting. The victim is emotionally invested in the gaslighter and in their relationship. Further, the gaslighter will typically be emotionally invested in the victim as well. In Sect. 2 I followed Sullivan-Bissett and Coltheart in treating confabulation as an ill-grounded explanation of an apparent gap in understanding where the ill-groundedness is due to motivational factors on the part of the confabulator, and to unusual or pathological features of the situation. Like gaslighting, confabulation thus has both epistemic and broadly normative motivational components. In this section I bring these two discussions together to argue that gaslighting will typically involve confabulation on the part of both the victim and the gaslighter. In the next section, I consider whether or not such confabulation is epistemically innocent.

First, I say that confabulation is *typically* involved in gaslighting, but this suggests "not always". Why not? Because I think that a certain kind of "pure", confabulation-free gaslighting is possible for both gaslighter and victim. Consider again Gregory from the 1944 film *Gas Light*. He seems fully aware of his intention to convince Paula she is losing her mind. He is fully aware that she is not really crazy and that he is methodically deceiving her and manipulating her by means of the trust she has in him. There is no "gap" between Gregory's actual intentions and the intentions he would willingly ascribe to himself (being the kind of man who would engage in this sort of deception and manipulation does not bother him), and no incongruence between how he thinks Paula is and how she really is (he is aware she is not crazy). Gregory has no need for confabulation, and none seems to be going on. Consider then his victim.

Paula clearly confronts the epistemic challenge I identified above: she must decide whether to trust Gregory's word and interpretation of what is going on or to trust her own cognitive faculties. Since Gregory is challenging her grasp of things in a rather comprehensive way, she must choose. Yet all appearances indicate that Gregory is trustworthy, and he not only asserts that Paula is losing her mind, but provides concrete instances where it seems clear that she is (the lights seem to her to be dim, but he assures her they are not; she does not remember moving or stealing something, but both Gregory and other witnesses (at his orchestration) testify that she has). There is nothing irrational about trusting another person's cognitive faculties over one's own in particular instances, and even about very important matters (Zagzebski 2012, Chap. 1).[15] There is no reason why a person should not even, under certain conditions, trust another person to tell them that they are cognitively malfunctioning in rather global ways, as when a close friend or trusted physician delicately informs someone that they may be suffering from Alzheimer's. Thus, there could be cases of gaslighting in which the gaslighter is so effective in providing his victim with good (though false) reasons to trust him, and in providing her with good (though false or fabricated) reasons not to trust herself, that the on-balance most reasonable thing for her to do is to accept his assessment. No gap in her evidence, so no bad abductive inference, so no confabulation is going on. Paula in the film *Gas Light* comes close to such a case.[16] Given the information available to her (as a result of her particularly thorough and rational gaslighter), she has reached an explanatorily adequate conclusion.

While Gregory of the *Gas Light* is possible, I think the typical gaslighter probably lacks full awareness or appreciation of the nature of his motivations, intentions, and behavior, and would strongly prefer *not* to think of himself as a deceptive manipulator. Returning to the idea, discussed in Sect. 1, that gaslighting is a kind of abuse, the foregoing characterization is consistent with the fact that participants in court-ordered batterer intervention programs typically display significant degrees of self-deception (Smith 2007; Vecina et al. 2016). Indeed, such individuals typically maintain that they have done nothing wrong, normalize their actions, present their punishment as unfair or as part of a system "rigged" against them, and present their partner (and victim) as equally responsible, ungrateful, or outright deserving of what happened (Smith 2007).

Transferring this model to the gaslighter, what seems likely is that (a) he will typically be unaware of or fail to fully grasp the nature of the motivations and intentions

[15] A near-sighted person without corrective lenses may well trust a normal-sighted person to help them navigate the park. Patients typically trust doctors to be better judges about health and disease.

[16] I leave to the side here the paradoxical nature of someone *rationally* deciding that their cognitive faculties are globally *untrustworthy*. After all, since it is their own faculties that ultimately get them to this conclusion, shouldn't they immediately distrust it? I think this is an interesting and significant issue and hope to make sense of it in future work. For the moment, to alleviate the pressure it puts on my description of Paula, I think that it is possible for a person to reasonably conclude that they are untrustworthy in a vast majority of cases or about the most important cases, even while retaining some trust in their ability to draw and accept the consequences of this very conclusion. All the gaslighter really needs, in order to *feel* that he has achieved his goal, is that his victim view herself as cognitively deficient and lacking a clear grasp of the situation *relative to him*. This is still a significant loss in epistemic self-trust and an act of subordination on the part of the victim, even if it does not amount to the total undermining that the gaslighter aims for.

guiding his behavior and (b) will not want to think of himself as the kind of manipulative deceiver that would engage in gaslighting. Given a gaslighter who satisfies these two conditions, confabulation is likely. First, it will be hard for the gaslighter to persistently ignore the fact that his behaviors toward the victim are atypical, condescending, or cruel.[17] Since he is committed to (b), this introduces a gap into his understanding that must be explained. That the victim in fact deserves this treatment because she is deeply wrong or mistaken about the situation would be a natural, if ill-grounded, explanation for the gaslighter to give, and hence confabulation. If, in addition, there are good reasons for the gaslighter to think that his victim is not so hopelessly out of touch with reality as his behavior suggests (as I take it there typically will be), then he confronts another gap in his understanding which will, in turn, require more confabulation.[18] Such a gaslighter is motivated to confabulate by two sets of desires: his desire not to think of himself as a deceptive manipulator, and his desire to have his victim accept his way of viewing things absolutely and without objection. These two desires conflict, so it would take a truly unaware gaslighter to never notice gaps or inconsistencies in his behavior toward his partner (victim), hence in the typical case such an individual will be driven to confabulate in one way or another. I don't insist that *all* gaslighters will have to confabulate in this way, but I think that many will and do, and this will be enough for the points I want to make below.

Concerning the role of confabulation in the victim's response to gaslighting, I think that most cases, and nearly all cases of successful gaslighting, will involve confabulation on the part of the victim. While it is possible that the best explanation for her available evidence really is that her own faculties cannot be trusted (as in the case of Gregory and Paula perhaps), this seems unlikely. In the typical case, the victim of gaslighting will have to rationalize or explain away some of the gaslighter's behaviors, at least early on. His comments, criticisms, and demands will strike her, rightly so, as a bit odd, out of place, or inappropriate. Perhaps, given her total evidence concerning him, it will even be reasonable at the beginning to ascribe such behavior to stress or tiredness on his part, or to some inadvertent and momentary failing on her own. However, as his gaslighting behavior persists and she finds herself under more and more emotional and psychological duress when she interacts with him, the growing evidence against his trustworthiness will call out for an explanation (she finds herself confronted with a "gap" in her understanding), and after a certain point this explanation will have to be confabulatory: an explanation that is *not* consistent with the victim's total available evidence, but adopted as a result of her strong desire e.g. to maintain the relationship with the gaslighter. Under these conditions, when the gaslighter is successful, he is so in part in virtue of the victim's having told herself a confabulatory story according to which he remains trustworthy while the trustworthiness of her own cognitive faculties or agency are downgraded in her own estimation. This is arguably the case of Pat from *Pat and Mike*. Unlike Paula, Pat does not find herself confronted with an intricate web of fabricated but convincing evidence for the conclusion that she shouldn't trust herself but should trust Collier. On the contrary, the gaslighting is rather transparent. Pat's total available evidence doesn't favor trust in Collier's sincerity or good judgment over trust in herself, so if she goes along with him (by actually believing what he says) she must tell herself an ill-grounded story about how he is right.

There is also empirical support for the idea that victims of abuse will rationalize or confabulate. It is relatively well-established that victims of childhood abuse develop strategies to isolate, minimize, or internalize (blame themselves for) abusive behavior on the part of someone they trust, and that this affects their ability to identify or label experiences *as* abuse (Goldsmith and Freyd 2005). These same strategies, specifically normalization of abuse and blaming one's self for the abuse, are found in adult victims of verbal abuse and domestic violence (Hannem et al. 2015). Victims of verbal abuse tend to normalize their verbal abuser's behavior and view themselves as at least partially responsible for it, explicitly calling it "abusive" only in retrospect after events have escalated resulting in more serious (and obvious) abuse. It is not hard to see how there would be parallels to this for typical victims of gaslighting. While the victim's best available evidence concerning her gaslighter may be that it is he, and not she, that should not be trusted (something she can't quite name is wrong with how he is treating her), her strong attachment to the gaslighter and fear of potential social stigma, coupled with the challenging nature of passing judgments of this sort to begin with, provide her with ample motivation to come up with an alternative explanation for the situation, one according to which the gaslighter's behavior is normal and it is she, not he, who is in some way at fault. It is in this sense that victims of gaslighting can and, I think, typically will confabulate.

[17] See Holroyd (2015), especially 514, and 519–521 for a discussion of how even a subject who is not fully aware of his precise intentions and motivations might nevertheless be held responsible for failing to identify morally relevant features of his own behavior and actions.

[18] Presumably a minimization of these reasons, or a reinterpretation of them that is consistent with his gaslighting behavior and view of the victim.

4 Epistemic Innocence

Recent research by Lisa Bortolotti and others has begun challenging the assumption that defective cognitive processes such as delusions and confabulations are blameworthy or categorially epistemically bad (Bortolotti and Cox 2009; Bortolotti 2015; Sullivan-Bissett 2015; Bortolotti and Sullivan-Bissett 2018). First, it is not clear that individuals who have confabulatory or delusional beliefs are responsible or should be blamed for this. For example, pathological confabulation seems to occur as a result of factors of which the confabulator is unware and that he is unable to access. At least some cases of non-pathological confabulation are similar.[19] Second, though not ideal, it turns out that processes such as confabulation and delusion *can* sometimes confer epistemic benefits not available by other means (Bortolotti and Cox 2009; Bortolotti 2015). This is significant because it suggests that merely condemning such processes, or discouraging them at all costs, may do more harm than good. This point, in turn, may have consequences for the correct way of treating patients who confabulate or are delusional in clinical (and perhaps non-clinical) contexts (Bortolotti and Sullivan-Bissett 2018).

When a cognitive process, such as confabulation, is indeed both the *best or only option available to the subject* and *confers some epistemic benefit rather than none*, that process is *epistemically innocent*. Bortolotti models epistemic innocence on the innocence excuse in law (2015, p. 495). A normally illegal action might be excused if the subject did not intend to perform the action, or if she took the action knowingly but in a context where taking it brought about some good or prevented some harm that no alternative available action could have. By analogy with this, a cognitive process is epistemically innocent if it meets two conditions:

> Epistemic Benefit: The delusional belief confers a significant epistemic benefit to an agent at the time of its adoption.
> No Alternatives: Other beliefs that would confer the same benefit are not available to that agent at that time. (Bortolotti 2015, p. 496)

If a cognitive process and the beliefs it produces are epistemically innocent in this sense then, Bortolotti argues, we should at the very least give it a second look in order to understand what is epistemically valuable about it, and how that value might be recovered and optimized even in the non-ideal situations where such cognitions occur.

4.1 Confabulation as Epistemically Innocent

Sullivan-Bissett has argued that confabulation, in cases such as her case of Roger and his confabulation due to implicit bias (discussed above), can be epistemically innocent. This is so first because Roger has no better alternative available to him at the moment when he must explain why he believes that Katie has a worse CV than the other candidates. While we may all have a general duty to become aware of and control for our implicit biases, Roger has not done this, so at the moment he confabulates concerning his belief he is unaware of the role of his implicit bias in causing it. Thus, while there are alternative and better explanations for Roger's total evidence (and thus alternative beliefs that would be more reasonable or better grounded for him), his implicit bias renders these *unavailable* to him at the moment he confabulates.

Second, Sullivan-Bissett argues that confabulation such as Roger's is epistemically innocent because it provides two types of epistemic benefit. First, in confabulating, a subject is at least forced to think about and provide explanations for his beliefs or actions in an explicit way, and doing this has a general tendency to lead agents to more true beliefs in the long run (p. 555). Insofar as confabulation is an occasion for reflection, and insofar as it forces the subject to confront the "gap" in his evidence or beliefs, this seems like a positive from an epistemic standpoint.[20]

The second epistemic benefit is that confabulation preserves coherence in a subject's beliefs and conception of themselves (Sullivan-Bissett (2015) p. 556). Confabulators don't tell just any story, they tell a story that closes a gap in their beliefs, and they typically do so in a way that maintains consistency among their beliefs and in their conceptions of themselves. Sullivan-Bissett recognizes the point, already raised by Bortolotti and Cox (2009, p. 557), that maintaining such coherence of self-concept in the case of confabulation typically comes at the expense of truth, challenging the idea that it is epistemically beneficial. However, she points out that confabulation can also have benefits of a non-epistemic sort that themselves may sustain or support epistemic benefits *indirectly* (pp. 555, 557). While maintaining a coherent self-concept by means of confabulation may come at the price of truth, there is value, perhaps even independent value, in self-trust (Pasnau 2015). If having a coherent

[19] e.g. Subjects who confabulated reasons for choosing the object furthest to the right in Nisbett and Wilson's experiment can hardly be blamed for failing to cite a "cognitive bias for objects located near the right hand" as the reason for their selection.

[20] Imagine if Roger went on forming beliefs based on his implicit biases and was *never* challenged to explain himself. This would surely be epistemically bad, whereas having to confabulate is arguably epistemically good insofar as doing so, or doing so often, encourages reflection which itself is more likely, all things being equal, to get the subject (Roger) to epistemically better beliefs in the future.

self-concept is the precondition for such self-trust, and so for the possibility of engaging in meaningful epistemic practices at all, then confabulation for this purpose at least keeps the confabulator "in the epistemic game".

4.2 Confabulation, Gaslighting, and Epistemic Innocence

I think Sullivan-Bissett is right that some cases of confabulation are epistemically innocent. However, considering the role of confabulation in gaslighting can help tighten up in helpful ways what it means for a practice to be *epistemically beneficial*. I've argued above that gaslighting will typically involve confabulation on the part of both gaslighter and victim. Here I will explain why such confabulation, even if the best alternative available, is not only not epistemically beneficial, but positively epistemically harmful. I don't take the fact that confabulation in gaslighting fails to be epistemically innocent to be a counter-example to Sullivan-Bissett's point about its epistemically innocent role in confabulation resulting from implicit bias. Sullivan-Bissett follows the literature on epistemic innocence in claiming that confabulation *can* be beneficial under certain circumstances, not that it *always* is (Bortolotti and Cox 2009, p. 557; Bortolotti and Sullivan-Bissett 2018). What the case of gaslighting does illustrate, however, is that whether or not a defective process such as confabulation will confer the cognitive benefits that the epistemic innocence literature typically identifies—namely (i) a greater likelihood of continued reflection and inquiry and (ii) maintenance of a coherent and efficacious self-concept which will, in turn, underwrite further inquiry—will depend in very specific ways on the context in which such faulty cognitions are deployed. While the literature thus far has used the key conditions of epistemic innocence to determine when it *does* apply to specific faulty cognitions (confabulation, delusion, clinical memory distortions, etc.), understanding more precisely when it *does not* apply is equally important. At least if the goal of developing the notion of epistemic innocence is, in part, to deploy it to positive effect in social and clinical contexts.[21]

That being said, why does the confabulation involved in typical forms of gaslighting fail to be epistemically innocent? First, it is plausible that, in many cases of gaslighting, no alternative and less faulty cognitions are *available* to either the perpetrator or the victim of gaslighting than the confabulations that they engage in.[22] Thus it probably isn't the "no alternatives" condition that confabulation in gaslighting violates. Sullivan-Bissett distinguishes between three types of availability: a cognition is "strictly unavailable" if it is "based on information that is opaque to introspection, or otherwise irretrievable"; a cognition is "motivationally unavailable" if it is inhibited or cannot be accessed due to motivational factors; and a cognition is "explanatorily unavailable" if it is, from the subject's point of view, so implausible that she is not even willing to entertain it (Sullivan-Bissett 2015, p. 554). She stresses the importance of taking particular factors of individual cases into account, but seems willing to consider the "no alternatives" condition of epistemic innocence satisfied as a result of any of these three types of unavailability, so long as the subject's confabulation is sincere and her confidence in it is sufficiently high to preclude her seriously doubting it (2015, pp. 557–558).

For the confabulating gaslighter and his victim, the unavailability at issue will most likely be either motivational or explanatory, or perhaps some mix of these two. If the strength of their motivations is sufficient to preclude serious consideration of alternatives to their confabulated beliefs, then the confabulation at issue satisfies the "no alternatives" condition. Similarly, if the degree to which alternative explanations are implausible to either the gaslighter or his victim is sufficiently great as to preclude serious consideration of them, then the confabulation at issue satisfies the "no alternatives" condition also. I do not think this will always be so. However, the type of gaslighter who is likely to confabulate is one who simultaneously has and acts on the motives and intention to gaslight, but is not explicitly aware of this fact, even while possessing a strong desire not to view himself as a manipulative deceiver. It is not difficult to imagine such an individual being sufficiently strongly motivated in both of these ways as to both confabulate about himself and his victim, and be unwaveringly sure that his confabulation is correct [as the self-deceiving abusers discussed above seem clearly to be (Sect. 3), and (Smith 2007)].

As for the confabulating victim of gaslighting, her confabulation seems to leave alternative cognitions both motivationally and explanatorily unavailable. The victim has trust in the gaslighter and also, typically, a strong personal investment. Her confabulated hypothesis is that the gaslighter's seemingly unusual and inappropriate behavior is, as he himself tells her, not actually unusual but is rather occasioned by shortcomings or failures in herself that she is unable to appreciate. While the victim may initially be unsure about her confabulation in this case,

[21] A project that seems to be quite productively afoot in (Bortolotti and Sullivan-Bissett 2018).

[22] It is not crucial to the point I want to make that alternative cognitions be unavailable to confabulators in the context of gaslighting. If such alternative cognitions *are* available, then their confabulations are obviously *not* epistemically innocent by this token alone. However, as I argue, I do think that alternative cognitions are often unavailable in this context, and I think that this being the case is better, as it places the focus squarely on the question of whether or not these candidates for epistemic innocence really are *epistemically beneficial*.

she has strong motivations to accept it and, once she does give it credence, she actually undermines her ability to trust future doubts concerning it. At this point she too seems to satisfy the "no alternatives" condition.

Concerning the "epistemic benefit" condition, however, the confabulation of gaslighter and of victim alike not only fail to confer epistemic benefits, but are positively harmful to them, and so clearly fail to satisfy this condition. The primary function of the gaslighter's confabulation is to enable his gaslighting activity, to underwrite and support his effort at gaslighting his victim. Further, it provides him with defeating evidence for the legitimacy of his victim's otherwise reasonable resistance to his actions. Confabulation *epistemically enables* the gaslighter, and if his victim begins to concede to his confabulation, even marginally, the confabulation becomes a self-fulfilling prophecy. While it is true that the confabulation allows the gaslighter to maintain a coherent and positive sense of himself as an epistemic agent, this seems to be precisely a case that raises doubts concerning the epistemic value of such "reflective agential coherence" taken by itself.

Similarly, the victim's confabulation gains her no epistemic benefit, as the confabulation she is most likely to engage in only aids and abets someone who is pushing her toward a total breakdown of epistemic self-trust. Whether the victim's confabulation simply rationalizes some instance of her gaslighter's behavior without undermining her trust in him, or involves conceding that some part of her experience, thought, or judgment concerning the situation is indeed defective, her confabulation either preserves the gaslighter's project unchallenged or actively assists it by providing her with the very beliefs that he desires her to have. Interestingly, the victim's confabulation does *not* allow her to maintain a coherent view of herself as an epistemic agent, though it does preserve coherence in other areas of her belief set and commitments by providing a rationale that renders consistent her trust in her gaslighter with his otherwise seemingly objectionable behavior. Thus, as in the case of the gaslighter, the fact that confabulation preserves some kind of coherence in the victim's belief set does not seem, in itself, to confer any epistemic benefit. The victim's confabulated story also does not seem likely to lead her to more true beliefs in the future. Indeed, it seems uniformly epistemically bad. After a certain point what may have begun as mere assertion and manipulation by the gaslighter takes on an evidential character and generates an epistemically poisonous feedback loop as both gaslighter and victim acquire self-perpetuating evidence for the truth of what they are each inclined to believe in any case.

4.3 Consequences for Epistemic Innocence and Confabulation

The upshot of this consideration of confabulation and gaslighting is that whether or not a defective cognitive practice counts as *epistemically beneficial* will depend greatly on the context, especially the social and interpersonal context in which it is deployed. In the case of gaslighting, the primary occasion for confabulation arises when the gaslighter and the victim interact with each other. However, this is a situation that is quite rigged, one where even protracted discussion is unlikely to result in more true beliefs, or even a tendency toward this. What this seems to show is that whether confabulation has positive epistemic benefits at all depends a great deal on the interpersonal context in which the confabulation is elicited. Merely having the occasion to provide confabulatory reasons for one's actions may have no tendency to produce more true beliefs in the long run if the context in which confabulation occurs also sustains and validates the explanations the confabulator offers. For example, one tacit racist might confabulate his reasons for not hiring an African American applicant for a position to another tacit racist only to find his confabulation confirmed and endorsed: it turns out they both think, contrary to the available evidence, that the candidate's CV was inferior to those of his non-African American competitors for the position. The epistemic benefits or harms of cognitions of the sort under consideration are particularly vulnerable to how others respond to them, both at a time and across time. Ideally, the reactions of others will minimize the epistemic harms of the cognition, while encouraging and assisting in the maximization of its epistemic benefits, but there is no guarantee of this.

Concerning maintaining a coherent sense of one's self and one's situation in the world by means of confabulation, the case of gaslighting seems to suggest that mere coherence is not sufficient for epistemic benefit, but only a coherent self-understanding that has some tendency toward the subsequent forming of true beliefs, or at least a tendency toward cultivating a greater openness to evidence and responsiveness to reasons on the part of the subject. Once again, if two tacit racists confabulate together and, in doing so, maintain a coherent sense of themselves and the world that leaves their racism entirely intact, it is not clear what if any epistemic benefit has been obtained. Not only do they sustain their false beliefs, but they also sustain the tendency to confabulate precisely in the way that is epistemically bad, as a way of avoiding or cutting off further reflection and questioning rather than engaging in it. I would propose, speculatively, that it should be possible to distinguish between coherent self-concepts that do confer some epistemic benefits, those that confer neither epistemic benefits nor harms, and those that confer epistemic harms, and to try to provide some characterization of the epistemic virtues and vices of each, and

of how these might be cultivated. The problem with confabulation in gaslighting is that, in the context it is deployed, it is not just epistemically neutral, but harmful in the types of self-concept that it sustains, and in the types of inquiry that it rules out. I imagine there may well be analogues of this in other cases.[23]

5 Conclusion

In the foregoing, I have developed two main lines of argument. The first is that prototypical cases of gaslighting involve confabulation on the part of both gaslighter and victim. Such confabulation is continuous with other instances of confabulation discussed in the literature, but provides a more detailed look at how confabulation might be deployed and sustained in the context of interpersonal relationships. Considering the role of gaslighting in interpersonal relationships, even deformed ones such as those involved in gaslighting, raises the possibility that the phenomenon of confabulation may be even more widespread than currently thought. Second, I have used the example of confabulation in gaslighting to motivate considering the role of contextual and psychological factors in determining when a defective cognitive process is epistemically innocent. In many prototypical cases of gaslighting, confabulations will be the only cognitions available, while also being epistemically harmful because they actually perpetuate the subject's epistemically harmful beliefs, practices, and dispositions. Seeing that and why this is so helps to more fully articulate the conditions under which a cognitive process such as confabulation is likely to be (or not be) epistemically beneficial, which is helpful both in clarifying the concepts of epistemic benefit (and so epistemic innocence) itself, and when considering possible applications of the notion of epistemic innocence in social and clinical contexts.

Acknowledgements For suggestions and conversation about gaslighting in all its forms I am especially grateful to Katherine Tullmann. An earlier version of this paper received helpful comments and criticism from all of the participants at the Grand Valley State Philosophy Summer Research Group, and from participants at the very stimulating Workshop on Confabulation and Epistemic Innocence organized by Elisabetta Lalumera. The final version of this essay is far better than it otherwise would have been thanks to generous and insightful comments from two anonymous referees for the journal.

Compliance with Ethical Standards

Conflict of interest The author declares that he has no conflict of interest.

Research Involving Human Participants or Animals The article does not contain any studies with human participants or animals performed by any of the authors.

References

Abramson K (2014) Turning up the lights on gaslighting. Philos Perspect 28(1):1–30

Bortolotti L (2015) The epistemic innocence of motivated delusions. Conscious Cogn 33:490–499

Bortolotti L, Cox RE (2009) Faultless' ignorance: strengths and limitations of epistemic definitions of confabulation. Conscious Cogn 18:953–965

Bortolotti L, Sullivan-Bissett E (2018) The epistemic innocence of clinical memory distortions. Mind Lang 33:263–279

Calef V, Weinshel EM (1981) Some clinical consequences of introjection: gaslighting. Psychoanal Q 50(1):44–66

Carpenter A (2018) Gaslighting America. Harper Collins, New York

Carruthers P (1996) Simulation and self-knowledge: a defense of theory-theory. In: Carruthers P, Smith PK (eds) Theories of theories of mind. Cambridge University Press, Cambridge

Carruthers P (2010) Introspection: divided and partly eliminated. Res 80(1):76–111

Christenson D (2009) Disagreement as evidence: the epistemology of controversy. Philos Compass 4(5):756–767

[23] Suggestions somewhat along these lines have recently been explored by Bortolotti and Sullivan-Bissett (2018). They consider the question of the epistemic innocence of clinical memory distortions (arguing that they can be epistemically innocent), and explicitly consider the resulting possibility that the epistemic innocence of clinical memory distortions might imply endorsing different treatment strategies for patients who suffer from memory impairments due to neuropsychological disorders. In particular, the thought is that perhaps it would be better for patients who report clearly distorted or confabulated memories to be supported and encouraged in their attempts to communicate and retrieve information about the past, however distorted, rather than outright contradicted and so potentially shut down. After all, if such individuals are regularly shut down when they speak about their memories, they may indeed shut down and become less social, so losing out on one important way in which to continue checking and forming new beliefs. Maintaining sociality is thus an epistemic benefit of patients' formulating and sharing such memory distortions (increased sociality is not the *only* epistemic benefit of clinical memory distortions identified by Bortolotti and Sullivan-Bissett, it is just the one most pertinent to my discussion here). On the other hand, if such distortions are allowed to stand essentially unchallenged, then it seems that maintaining or increasing sociality might undermine the epistemic good (ongoing inquiry and the possibility of checking and correcting one's beliefs with others) that it is in part supposed to sustain. Bortolotti and Sullivan-Bissett's discussion is much more wide-ranging than just this point, and they do not explicitly endorse the therapeutic consequences that they consider. What I think their discussion does show, however, is the complex relationship between non-optimal cognitions, their epistemic benefits (and harms), and the social or clinical context in which these are deployed. The same kinds of relationships that I have explored here in the context of gaslighting and confabulation.

Coltheart M (2017) Confabulation and conversation. Cortex 87:62–68
DeWitt R (2010) Worldviews: an introduction to the history and philosophy of science. Blackwell, New York
Galileo G (1632/1989) Dialogues concerning the two chief world systems. In: Matthews MR (ed) The scientific background to modern philosophy. Hackett, Indianapolis
Goldmsith RE, Freyd JJ (2005) Awareness for emotional abuse. J Emot Abuse 5(1):95–123
Gopnik A (2000) Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In: Keil JC, Wilson RA (eds) Explanation and cognition. MIT Press, Cambridge
Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgment. Psychol Rev 108:814–834
Hamilton P (1938/2015) Gaslight. Samuel French Ltd, New York
Hannem S, Langan D, Stewart C (2015) Every couple has their fights… Stigma and subjective narratives of verbal violence. Deviant Behav 36(5):388–404
Hirstein W (2005) Brain fiction: self-deception and the riddle of confabulation. MIT Press, Cambridge
Hirstein W (2009a) Confabulation: views from neuroscience, psychiatry, psychology, and philosophy. Oxford University Press, Oxford
Hirstein W (2009b) Introduction. In: Hirstein W (ed) Confabulation: Views from neuroscience, psychiatry, psychology, and philosophy. Oxford University Press, Oxford
Holroyd J (2015) Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition special issue on Imperfect Cognitions* 511–522
Kopelman MD (2010) Varieties of confabulation and delusion. Cognitive Neuropsychiatry 15:14–37
Kuhn TS (1977) Objectivity, value judgment, and theory choice. In: Kuhn TS The essential tension. University of Chicago Press, Chicago
McKinnon C (2003) Knowing cognitive selves. In: DePaul M, Zagzebski L (eds) Intellectual virtue: perspectives from ethics and epistemology. Oxford University Press, Oxford
Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on memory processes. Psychol Rev 84:231–259
Pasnau R (2015) "Disagreement and the value of self-trust. Philos Stud 172:2315–2339
Smith ME (2007) Self-deception among men who are mandated to attend a batterer intervention program. Perspect Psychiatr Care 43(4):193–203
Stern R (2007) The gaslight effect. Random House, New York
Sullivan-Bissett E (2015) Implicit bias, confabulation, and epistemic innocence. Conscious Cogn 33:548–560
Vecina ML, Chacon F, Perez-Viejo JM (2016) Moral absolutism, self-deception, and moral self-concept in men who commit intimate partner violence: a comparative study with an opposite sample. Violence Against Women 22(1):3–16
Wagers SM (2015) Deconstructing the 'power and control motive': Moving beyond a unidimensional view of power in domestic violence. Partner Abuse 6(2):230–242
Zagzebski L (2012) Epistemic authority: a theory of trust, authority, and autonomy in belief. Oxford University Press, Oxford