1

The Powers That Bind: Doxastic Voluntarism and Epistemic Obligation

Neil Levy and Eric Mandelbaum

As the phrase is usually used, 'doxastic voluntarism' is the thesis that agents have the power to directly form beliefs for non-epistemic reasons. The thesis that we have such a power is an interesting one, one that is worthy of exploration on its own terms. However, it is often discussed because of its close connection to an even more interesting question: whether we have any epistemic obligations. The connection between the two is motivated by some version of the ought-implies-can principle; the thought is that we only have obligations to come to hold beliefs with particular contents if we have the power to form such beliefs.

In this paper, we argue for three theses: (1) we lack the power to form beliefs at will (i.e., directly); at very least, we lack the power to form at will beliefs *of the kind* that proponents of doxastic voluntarism have in mind;[1] but (2) we possess a *propensity* to form beliefs for non-epistemic reasons; and (3) these propensities—once we come to know we have them—entail that we have obligations similar to those we would have were doxastic voluntarism true. Specifically, we will argue that we have obligations to avoid triggering these propensities to form beliefs that are unwarranted or even immoral. We therefore issue a warning: if you read this paper, you will find yourself with more obligations at the end than you currently possess.[2]

[1] One of us believes that we lack the power to form beliefs at will while the other believes that we lack the power to form at will beliefs of the kind that proponents of doxastic voluntarism envisage. This is a dispute about the nature of beliefs, not about the nature of our powers.

[2] Indeed, it may already be too late: ceasing to read now might constitute what Smith (1983) calls a benighting act, by virtue of which you are culpable for your ignorance of your epistemic obligations.

1.1 Truth-Critical Deliberation and Voluntarism

If we had the power to directly form beliefs (of the kind that proponents of doxastic voluntarism envisage; from now on we drop the qualification except when it is under discussion) for non-epistemic reasons, we might be required to justify the beliefs we form in this way. The request for justification would be significantly different from the request we routinely make of one another: rather than asking what evidence we can cite in favour of the belief's being true, a request for justification might ask for non-epistemic—moral or prudential, say—reasons for holding that belief. This would be an additional requirement, additional, that is, to the requirement we are sometimes under to justify our beliefs by reference to facts concerning their likelihood of being true. It would also be a more demanding requirement, inasmuch as these acts of belief formation would be voluntary. Voluntary behavior is, *ceteris paribus*, behavior that is apt for blame and praise, whereas non-voluntary behavior probably isn't directly apt for blame or praise.[3]

It is widely agreed that we do not have the power directly to form just any beliefs. I cannot directly decide to believe that today is Wednesday, for instance. Call a belief that has no prior epistemic support an arbitrary belief. That *strong voluntarism* (Frankish 2007), the thesis that we have the power to directly form arbitrary beliefs, is false is more or less universally accepted. But *weak voluntarism*, the thesis that we have the power directly to form beliefs given certain epistemic conditions, is more controversial. We claim it too is false, and for precisely the same reasons as strong voluntarism. Strong voluntarism is false because forming an arbitrary belief with the content *p* requires that we simultaneously bring it about either that we forget that we have the belief only because we have decided, for non-epistemic reasons, to form such a belief, or that we change our view of the evidence so that we take the belief to be justified independent of our act of belief-formation. But we do not have the power to do either of these things directly. We cannot directly alter the contents of our memory at will, nor can we directly alter our view of the evidence at will. (No doubt we can take indirect means to alter either our memory or our view of the evidence; we might, for example, hit ourselves in the head with a brick after we form a belief hoping that the ensuing amnesia knocks out our memory of the belief formation process without altering the belief itself. Such routes are clearly not direct in the relevant sense.) It might be thought that it would be no harder to alter our view of the evidence than it would be to change our belief—our view of the evidence is just another belief, after all. But the fact that our view of the evidence is just another belief doesn't entail that we can alter it at will: altering this belief would require altering our memory that we have done so, or altering

[3] The claim that voluntariness is a necessary condition for blame is a controversial; one of us has defended the claim at great length in a number of places (e.g., Levy 2005; 2011). For opposing views, see Adams (1985) and Smith (2005).

our view of higher-order evidence, and so on. Because we cannot complete an infinite series of acts, we can't get ourselves to believe at will (Frankish 2007).

Frankish argues that the facts just mentioned entail that it is nomologically non-contingently true that strong voluntarism is false. We shall suggest that he is wrong in the following way: Frankish is right that strong voluntarism is false, but wrong in thinking that it is non-contingently false. It is a metaphysically contingent (though perhaps psychologically necessary) fact about us that we are unable directly to form beliefs for non-epistemic reasons. More importantly, perhaps, Frankish is also wrong when he asserts that there are no grounds for thinking that weak voluntarism is false.

Frankish believes that under certain conditions, we can directly bring ourselves to have a belief. These conditions are epistemic: we must have better evidence for the belief than for its negation. In other words, we can directly form a belief when, and only when, so doing takes us from fence-sitting to belief.[4] We do this, Frankish suggests, by formulating a policy of relying upon the content (i.e., taking the content to be true in a non-pro-tem fashion) in what he calls truth-critical deliberations, where deliberation is truth-critical when it relies upon premises the subject is disposed to accept in contexts in which truth is of central importance. (Frankish suggests that we can identify these premises with those premises we are apt to rely on in most contexts.) So doing just *is* forming the belief, so in doing this we directly bring it about that we have the relevant belief. We will not forget that we have the belief as a result of adopting the policy, nevertheless we will retain it because our view of the evidence *permits* us to believe that *p*.

We argue that these claims are false.[5] We can certainly adopt a policy of relying upon a claim in deliberation, but in so doing we will not bring it about that we possess all the dispositions constitutive of, or entailed by, a belief; not at once, at any rate.[6] Frankish appreciates the need for an act of belief formation to bring it about that the agent has the appropriate dispositions and intends his account to satisfy this condition. He believes that adopting a policy of taking *p* as a premise in truth-critical deliberation makes one disposed to believe that *p*.[7] But while it is possible that adopting such a policy will bring

[4] Ginet (2001) defends a similar position.

[5] One of us believes that though adopting a policy of relying upon a proposition does not bring about all the dispositions typically associated with a belief, nevertheless it does cause a state that deserves to be called a belief. This is because this author thinks that merely entertaining that *p* causes one to believe that *p*. On this view we, strictly speaking, do not form beliefs for reasons at all (this claim holds over perceptual beliefs but a bit more subtlety is needed for dealing with beliefs that are inferred as consequences from other beliefs). Since this view is, to put it mildly, not mainstream, we will ignore the view in the body of the paper so as to keep contact with mainstream usage, though we will occasionally note how adopting this view of belief would affect some of our claims.

[6] Of course adopting a policy might bring us sooner or later to have the correlative belief, but no one denies that we can alter our beliefs in this indirect manner.

[7] Using a premise in truth-critical deliberation is sometimes referred to as 'accepting' a premise—e.g., Alston (1996) and Bratman (1992). We shy away from this terminology since we think it causes confusion once the Gilbert framework is on the table.

it about that one has some of the dispositions characteristic of a belief, it will not bring about all of them all at once. A sufficient number of dispositions central to the dispositional stereotype associated with the belief will not immediately follow suit. In particular, the phenomenal dispositions (Schwitzgebel 2002)—the dispositions to have the appropriate affective responses—will not follow all at once. Someone who adopts a policy of taking *p* as a premise in truth-critical reasoning does not thereby cause themselves to be surprised if very soon after it is demonstrated to her that ~*p*. Such a deliberator will not exhibit any more surprise at such a demonstration than previously, when she was a fence-sitter and thought it somewhat likely that *p*.[8]

So adopting a policy of taking *p* as a premise in truth-critical deliberation does not directly bring it about that one has the correlative belief.[9] In fact, weak voluntarism fails for precisely the same reasons as strong: we will acquire all the dispositions associated with a belief that *p* (as opposed to the dispositions associated with thinking that *p* is more likely than ~*p*, which by hypothesis the agent has prior to adopting the policy) only if we simultaneously bring it about either that we forget we have the belief as a result of adopting a policy or that we alter our view of the evidence.

AQ: Is this still forthcoming, or is there a date of publication? – AAH OUP online suggests June 2014 (est.)

Frankish's mistake arises due to his apparent tendency to think of beliefs as all-or-nothing states. If belief was an all-or-nothing state, then it might be possible to move from non-belief to belief by way of adopting a policy, given that prior to adopting the policy one believed that the evidence was such as to make the belief more likely than not. But beliefs are not all-or-nothing states. Rather, they come in degrees (*pace* Holton forthcoming). This being the case, successfully bringing it about that one believes that *p* occurs just in case one has caused one's subjective probability that *p* to rise by some nontrivial amount. It is this that we cannot do, all at once, by behaving as Frankish recommends.

1.2 Basic and Non-Basic Actions

Though we think that Frankish is wrong to think that we can go from being fence-sitters to believers all at once, in the manner he recommends, we think that the facts that ensure that we do not have this power (and which also ensure—as Frankish

[8] If one were beholden to the view that entertaining causes belief, then one would have to deny that the phenomenal dispositions are in any sense constitutive of belief. One of the authors does so deny that any particular phenomenal states are even associated with, never mind constitutive of, belief.

[9] Two possible objections might be raised to this claim. First, one might object that if I take *p* as a premise and then use it to derive a conclusion that I have antecedent reason to believe is true (but didn't know it followed from *p*) that might cause me to raise my credence in *p*. But even if this were right, this wouldn't count as *directly* raising the credence in *p* merely by using it as a premise in truth-critical deliberation. Second, one might object that the mere activation of a thought raises its credence (á la Mandelbaum 2010) and the more one uses a premise in deliberation the more that premise will be activated. However, this sort of evidence applies to states that aren't the full-blown beliefs that Frankish discusses. Accepting this line of thought would be consistent with the non-mainstream view of belief that's been mentioned in the footnotes, and not the view that Frankish maintains.

recognizes—that we do not have the power to form any arbitrary beliefs) are only contingently true. Those facts, recall, are that we succeed in forming the belief that *p* only if we can simultaneously bring it about that we forget that we have formed the belief that *p* for non-epistemic reasons, or we can change our view of the evidence. We think it is a contingent fact that we cannot do these things and, therefore, a contingent fact that strong voluntarism is false.[10] An agent who could do one or both of these things is a conceptual, and perhaps even a genuine empirical, possibility.

As we are using the terms, people *directly* bring themselves to believe that *p* if they believe that *p* immediately upon performing some basic action, which they perform because they intend to bring it about that they believe that *p*. We maintain that in order to be successful, this basic act must bring it about that they forget how they brought about the belief, or alter their view of the evidence. As it happens, we do not know how this can be done: we have no idea what steps an agent might take to bring it about that they achieve these things. But we think it is conceptually, and perhaps even empirically, possible that there are steps that an agent can take that would bring it about that they acquire the power to believe at will.

A basic action is an action performed without any intermediaries. Raising one's hand is a basic action for most of us because we do not raise our hands by doing anything else; rather, we just raise our hands. Agents for whom raising one hand is not a basic action are actual: an agent suffering from paralysis of one arm, for instance, might only be able to raise the hand on that side by way of doing something else (grasping it with their other hand, perhaps). Now, the basic/non-basic distinction, so understood, is not the distinction between actions which are causally complex and those that are causally simple. Raising one's hand counts as a basic action even if neurally there are many stages involved (and there are). Rather, the basic/non-basic distinction is a distinction concerning how direct the action is for the agent: subpersonal complexity does not map onto personal directness. This fact entails that precisely the same action can go from non-basic to basic as the agent becomes more skilled at performing it.

Consider a recent example of how agents have performed an action by way of doing something else. Building on earlier work that showed that some patients diagnosed in a vegetative state were able to perform a task in which they could voluntarily imagine playing tennis or navigating a familiar environment (Owen et al. 2006), inasmuch as the neural activity they exhibited did not differ significantly from controls, Monti et al. (2010) were able to develop what was, in effect, an fMRI-based communication system, in which a patient, again apparently vegetative, was able to answer 'yes' or 'no' to questions by imagining playing tennis or imagining navigating a familiar environment. Responding to these questions was, for him, not performing a basic action. Rather, it was performing an action by way of performing another, an act of imagination, which was basic. However we believe that it is not merely possible but even quite

[10] It should be noted, however, that some people have denied that it is false at all. See, for instance, Steup (2000).

likely, given technological developments, that control over a communication device like this could become automated. Someone might eventually learn to give the correct responses so efficiently that for them it would be the basic action of < answering the question >. Similarly, we believe, someone might learn to control a prosthetic device using an EEG-based control system by a series of stages, beginning with discovering (say) that they can cause it to perform a desired movement by imagining a certain motor response, but ending with them moving it by performing the basic action of moving the device in the desired manner. When this occurred, a non-basic action would have become a basic action.

For the transformation of non-basic actions into basic to occur, the agent must learn to act with a high degree of efficiency and reliability. At the moment, the kind of indirect control that agents exercise over their beliefs,[11] is neither efficient nor reliable. Instead, it is very much a hit-or-miss affair. We change our beliefs in this indirect way[12] by the kinds of means that Pascal recommended to the person who wanted to bring about belief in God: associate with believers, immerse yourself in religious writings, try to think and act like a believer; eventually, perhaps via the mechanisms of cognitive dissonance reduction, you may find yourself with the correlative belief. If and when that happens, you will not forget that you have the belief via a process of self-manipulation, but you will find yourself with a different take on the evidence. From your perspective, it will seem to you that you have manipulated yourself into holding a belief that is independently warranted.

Now, if some day we hit upon a method to reliably and efficiently induce these changes in ourselves, it might become possible to automate the process. For someone who automates the process, they will be able to perform the basic action of changing at least some of their beliefs. Such a person would be like Jonathan Bennett's (1990) Credamites, who can will themselves to have a belief, except we think it is more realistic to suppose that agents who were much like us could perform such a basic action of willing belief by bringing themselves to have a different view of the evidence, rather than by forgetting how they brought the belief about. We say that because as a matter of fact real agents can and do indirectly induce beliefs in themselves, in the way recommended by Pascal, but in these actual cases the trick is performed by changing the agent's view of the evidence: it is this trick that is available to be automated.

1.3 Belief Acquisition on the Cheap

Though we do not believe that agents have the kinds of powers needed for doxastic voluntarism to be true, we do believe that we can—and do—form beliefs for non-epistemic reasons. In this section, we want to delve into the psychological

[11] At least in cases in which they can't make a belief true or false by acting directly on the conditions that make it true or false; say, making the belief that the light is off true by turning the light off (Feldman 2001).

[12] Setting aside science fiction cases involving direct stimulation of the brain or memory-erasing pills.

literature concerning how people actually form beliefs. After all, doxastic voluntarism is an empirical claim: it's a claim about whether people actually can directly form beliefs for non-epistemic reasons. Even though some theorists in the literature attempt to deal with such a claim through conceptual analysis alone,[13] we think it best to interweave such analysis with empirical findings. Thus we will now turn our attention to the literature on irrational and arational belief formation. In doing so, we will illuminate what powers and propensities to form beliefs for non-epistemic reasons human beings actually have.

As a warm up, consider some findings from what we might term 'the irrational belief formation' literature. It has long been noted that motivated reasoning can affect one's interpretation of evidence. When motivated reasoning does so affect one's belief acquisition capacities, the result is a belief that is formed for reasons and hence is capable of being assessed both psychologically and epistemically. However, because the belief formation processes here are motivated by non-epistemic values and goals, the end result is generally less than epistemically respectable. Perhaps the most famous study stemming from this tradition is Hastorf and Cantril's (1954) 'They Saw a Game: A Case Study'. Hastorf and Cantril showed Princeton and Dartmouth students a tape of a (then recent) very rough American football game. Both sets of students watched the same film yet on average Princeton students saw Dartmouth players commit twice as many infractions as the Dartmouth students saw. Moreover, perceptions of the severity of the infractions also greatly differed between the two groups. Hastorf and Cantril's venerable finding is now part of the common background knowledge on belief formation: what people want to see greatly affects their interpretation of the events they perceive. The irrationality inherent in these findings is that the students would or could not form impartial perceptual beliefs.

This type of means-end sifting through the evidence is typical of other effects in the psychological canon that can also be filed under 'motivated reasoning' such as the confirmation bias. The confirmation bias can be found in different guises. For example sometimes it's seen as a form of biased assimilation, sometimes as a biased information search.[14] But on either reading, the phenomenon looks to be one where people form beliefs through a biased strategy with the end of making people reaffirm their already held beliefs as opposed to objectively viewing new evidence. In Lord et al. (1979), subjects were shown mixed evidence about capital punishment. The evidence was completely equivocal—for instance, one piece of evidence consisted of data that capital punishment had positive effects on both past and future murder rates, and another piece consisted of data that pointed to the opposite conclusion, and both had equal evidentiary value. People who had antecedently believed in capital punishment claimed that the evidence presented against capital punishment had little probative

[13] See for example Hieronymi (2006) and Setiya (2008).

[14] Of course, sometimes it's just the name of a positive test search (such as in Klayman and Ha 1987); that use of the phrase is orthogonal to our purposes and should be set aside.

value, whereas the anti-capital punishment folks claimed that the evidence presented in favor of capital punishment was unpersuasive. Furthermore, both groups ended up with more polarized attitudes after being exposed to evidence that ran contrary to their opinions: both groups ended up believing more in their antecedently held views after encountering evidence that was problematic to their belief system!

Such data comes as no surprise to those who are familiar with the literature on cognitive dissonance. Displaying the effects of selective exposure to information is one of the core tricks in dissonance theory. The selective exposure effects show that people do not sift through evidence in an objective fact-seeking way; rather, people attempt to search for information which confirms what they already believe while avoiding information that might contradict what they believe. For example, Brock and Balloun (1967) played messages for subjects that warned of the ill effects of smoking, particularly the connection between smoking and cancer. These messages were interlaced with heavy static, which could be shut off by pushing an 'anti-static' button. The non-smokers reliably pushed the anti-static button more than the smokers. However, when the message was changed to one that disputed the link between cancer and smoking, the smokers reliably pushed the anti-static button more than the non-smokers. The same moral held true for churchgoers and atheists when they were asked to listen to a message that attacked Christianity: the churchgoers were happy to endure the static that made the anti-Christian message unintelligible. Of course, these are just a few examples from a deluge of work showing people's relative receptivity to information that confirms their antecedent belief and their hostility to and avoidance of counter-attitudinal information.

All of the effects canvassed so far can be understood as somewhat irrational effects on belief fixation.[15] They are irrational because they (a) aren't normatively respectable and (b) are explicable at the psychological level, a level of explanation where speaking of rational and irrational inferences and tendencies makes sense. However, there is also evidence about belief fixation that operates below the psychological level, evidence which is ground zero for theories that want to talk about descriptively adequate models of belief fixation.

Certain forms of belief acquisition cannot be given the honorific of 'rational' or 'irrational'; in order for something to be irrational it has to have a certain type of etiology. Let's return to our aforementioned friend, the brick. Suppose you get hit in the head with a brick and the force of the brick causes you to believe, for no reason at all, that the universe has ten planets. Now no doubt, this would not be a particularly well-justified belief, but it would be odd to condemn you for your mode of belief acquisition. After all, it isn't bad reasoning that led you to this belief. Instead, you formed this belief in a merely brute causal way. It is this type of causal process, brute causal incursions from beneath the psychological level causing certain beliefs, that we will term *a*rational.

[15] Lexicographic note: we use 'belief acquisition' and 'belief fixation' synonymously.

Arational belief formation is frightening because it is, on the face of it, seemingly impossible to counteract psychologically and very difficult to counteract at all. But more frightening still is the ubiquity of arationally caused doxastic—belief-like—states. In a series of fascinating studies Dan Gilbert and colleagues have accumulated evidence showing that people acquire belief-like states in a brute causal way. In particular, the work of Gilbert et al. appears to show that we are disposed to go on to form these states corresponding to any arbitrary proposition we happen to entertain.

The basic arational paradigm exploits asymmetries in people's memory of truths and falsehoods. In a typical experiment, participants are asked to participate in a learning task while they are intermittently placed under cognitive load and are then tested about what they learned. The recurrent finding is that when the learning occurs under even slight cognitive load, people tend to misremember statements that they learned were false as true, but do not tend to misremember true statements as false. An example should illuminate the situation. In one telling experiment participants were asked to learn nonsense word meanings. They watched a computer screen where sentences of the form 'An X is a Y' appeared, in which the 'X' was a nonsense word and the 'Y' was a word in English (e.g., 'A suffa is a cloud', from Gilbert et al. (1990)). Right after participants read a sentence the screen flashed either the word 'true' or the word 'false', indicating whether the previous statement was accurate or not. Participants were also told to be on guard for a tone that would occur; the tone would occasionally sound and when it did the participants were to push a button as soon as possible. The tone was introduced in order to induce cognitive load. During the critical trials, participants read six true and six false claims. While reading four of these claims (two true, two false), the participants were interrupted by the tone (these were the critical trials, since load was occurring). At the end of the trials the sentences were then turned into questions (e.g., 'Is a suffa a cloud?') which the participants then answered. The added cognitive load did not effect the true statements: participants reliably encoded true statements as true. However, the load did significantly affect performance on false statements: false statements were consistently incorrectly encoded as true.

Lest one think that the asymmetry between remembered truths and falsehoods holds just over 'mere memory', perhaps one more example would help to show how this acquired information is used in a belief-like manner. In Gilbert et al. (1993) participants were asked to watch a video screen with two crawling lines of text on it, one on top of the other. The top scroll contained text reports of two unrelated crime incidents. Participants were told that they would read both true and false details about the incidents, true statements appearing in black, false statements appearing in red. The bottom crawl did not contain any text, but instead had digits that slowly moved across the screen. Half the participants—the unburdened participants—were told to ignore these digits whereas half—the burdened participants—were told to peruse the digit crawl and to push a button anytime the number 5 appeared.

At the conclusion of the video, participants were asked to recommend a prison sentence for the offenses, ranging from zero to twenty years, and they were also asked

to assess the criminal's personality. In particular, participants were queried as to how much they liked him, how dangerous he was, and how much counseling would help him. The false statements the participants read during the first phase of the experiment either exacerbated or mitigated the severity of the crime. The participants in the burdened condition were significantly more likely to be persuaded by the false information. The participants in the unburdened condition recommended a sentence of six years when the false information was extenuating and seven when it was exacerbating—not a significant difference—whereas their burdened counterparts recommended five years in jail in the extenuating condition and eleven years in jail in the exacerbating, which is a statistically significant difference. Significant differences were also found across the board when looking at the defendant's likeability, benefit from counseling, and dangerousness. Thus, it appears that the falsehoods became integrated with the participants' beliefs and affected a robust range of their responses. If they were not yet beliefs—we doubt that Gilbert's subjects would have had all the dispositions associated with the correlative belief[16] —they were clearly on the way to becoming full-blown beliefs. They certainly affected their beliefs proper, perhaps by biasing the manner in which they processed information.

The propositions that the participants encountered while under load rippled through their cognitive system. In the first part of the study the participants not only processed the lies fed to them, but they made—presumably unconscious—inferences from those states which then informed their judgments concerning the duration of the sentence and the character's likeability. This is quite interesting because it shows that the false information that is acquired acts like beliefs in a hitherto unseen way: the information is informationally promiscuous, a hallmark of beliefs. Informational promiscuity has been previously suggested as a criterion for separating beliefs from other belief-like, sub-doxastic states, such as intramodular representational states, e.g., the representations inside one's language module (see Stich 1978). The attitudes the participants formed infiltrated and interacted with (presumably some subset of) their web of belief in order to produce the behavior the experiment detected.[17]

The asymmetries we have been discussing, ones between encountering truths and falsehoods while distracted, can be seen throughout the literature: a person put under cognitive load is apt to remember statements that they are told are false as true but not statements they are told are true as false. The experiments above displays that affirming a proposition (i.e., remembering the proposition as true)[18] comes much easier than

[16] In particular, we doubt that they would have asserted the belief. Of course, if you think that entertaining that *p* causes one to believe that *p* then you will sever the connection between belief and assertion. So such a theorist would think that although the subjects wouldn't necessarily assert that they believe the false propositions they encountered, they'd still act as if they believed them, as we see in the aforementioned study.

[17] Note that as far as this use of inferential promiscuity is concerned, what matters is that the information was available to a whole host of other processes and not that people were running honest-to-god inferences on the information.

[18] Here 'affirm' and cognates should not be read as entailing consciousness of the content let alone intentional or effortful mental action.

rejecting a proposition (i.e., remembering the proposition as false). Affirming is easier because it is a passive process, whereas rejecting is an active one: our cognitive architecture is set up to immediately affirm propositions as true. To go further and reject those propositions takes mental effort that is not necessary for the affirmation of a proposition. That is why something like belief fixation, operationalized above as the learning of sentences, can occur under load, but the rejection of a proposition—operationalized above as remembering that something is false—stalls when one is under cognitive load. The added cognitive load helps to shortcut the active rejection, but does not interfere with passive affirmation because the passive process is automatic and load does not affect a reflex. Compare how counting backwards from one hundred by increments of five would affect *seeing* a crossword puzzle versus *completing* the puzzle. The former will not be affected while the latter will be greatly affected. Rejecting a proposition is more like thinking than seeing, while affirming is more like seeing than thinking.

The observed asymmetry can be explained if we assume that when propositions are initially processed they are encoded as true by default and can only subsequently be marked as false. Evidence for this view comes from a disparate array of sources and because of space constraints we couldn't possibly canvass all of them (though see Mandelbaum (2010) for a painstakingly thorough review). However, before we leave the topic, we will describe one other experimental paradigm that speaks in favor of the mere-entertaining-causes-affirmation view. Instead of looking at acquisition of propositions that are personally meaningless we will now move our focus to forming beliefs about our own skills. To do so, we turn our attention to studies of belief perseverance in the face of experimental debriefings.

In Ross et al. (1975) experimenters asked participants to read a collection of suicide notes and to sort the real ones from the fakes. Participants encountered twenty-five pairs of notes and were told that one note from each pair was a real note, the other a fake (in fact, all were fakes). After seeing each pair participants would judge which note was real and which fake and were then given feedback on their performance. After receiving the feedback the participants were partially debriefed. During the debriefing the participants were told that all the feedback they received were fictitious, it being arbitrarily determined beforehand regardless of the participants' responses. After the debriefing the participants were asked to estimate both how many times they actually answered correctly and how many correct answers an average person would give. Interestingly, the information in the debriefing session did not affect participants' opinions about their ability: if the participant originally received positive false feedback (e.g., twenty-four out of twenty-five correct), they believed that they were better than average at the task, and if they received negative false feedback (e.g., seven out of twenty-five correct), they believed they were worse than average at picking out real suicide notes from fake ones.

The aforementioned experiment is not generally taken to illuminate anything about belief acquisition per se. It seems that the participants formed their beliefs in a

reasonable way, based on the experimental feedback. Once they are told that the feedback was non-veridical they may just have had trouble updating their beliefs. Perhaps beliefs are 'sticky', in that once one has a belief, that belief is hard to relinquish. If so, then the debriefing effect wouldn't tell us anything about belief acquisition *per se*, but rather belief perseverance.

But what happens if the subjects are briefed before they take part in the study and receive false feedback? (Call such a technique 'prebriefing'.) What if before sorting the notes they are told that the feedback they are about to receive is bogus? It turns out that prebriefing the participants has the same effects on subjects' beliefs as false feedback.

Wegner et al. (1985) replicated the Ross study except the participants were told *prior* to the task that the feedback would be dubious. Even after the explicit prebriefing the participants continued to behave as if the feedback was veridical. They were unable to reject the feedback they received, even though they knew it was bogus. These perseverance effects are easily explicable if we assume that the knowledge of the feedback persists because the participants automatically affirm the feedback when they hear it, even though they know the feedback is false. Since they are engaged in a relatively fast-paced experiment, the participants lack the mental energy to override the false claims.[19]

Although we have discussed only a few of the results from the empirical literature on belief acquisition, we think it's wise to conclude that our belief-fixating faculties have been set up in the following way: we are designed to initially affirm any propositions that we happen to think about. In the absence of the time and resources to reflect on these affirmations we will acquire belief-like doxastic states, and—soon enough—tend to acquire the correlative belief. Our cognitive architecture is set to automatically lead to affirmation of the propositions we happen to token. Thus, belief-like states come cheap: whatever we happen to encounter, we are disposed to affirm, and eventually to believe. Only after the initial acquisition of the proposition can we go back and reject the information we have acquired. Rejection differs from the initial affirmation in that the rejection is neither automatic nor ballistic; rather, rejecting a proposition is an active mental endeavor.

One's level of education and intelligence—whatever that exactly is, if anything—does not affect one's proposition-affirming faculties in the first instance. Instead, all of us are set up with dispositions to acquire beliefs in brute causal ways. Education and

[19] There is an interesting question whether subjects in both kinds of belief perseveration paradigms form full-blown beliefs of the type Frankish discusses, or whether prebriefing brings about a doxastic state with a narrower set of dispositions than the state that is acquired in the debriefing paradigm. One might think that the difference in these paradigms is that in the debriefing paradigm belief formation is evidence-based, whereas in the prebriefing one it is not. Perhaps this difference explains why the first results in full-strength beliefs. However, it's unclear in what manner the states acquired in the prebriefing studies differ from the debriefing ones. If prebriefing also results in full-strength beliefs, then this explanation looks more strained (though perhaps even in this condition beliefs are formed *because* something that looks evidence-like is presented). The more one is apt to see the same states acquired in both paradigms, the better the non-standard entertaining-is-believing line should look to one.

intelligence are tools that can help the rejecting process, for example by giving us more motivation and greater levels of concentration needed to reject certain propositions,[20] but do not affect the initial process, for the process works below the psychological level as it were (and since it works below the psychological level, the process appears to be arational). Evolution has conspired to make us initially gullible, a decent strategy for creatures like us who have more or less veridical perceptual faculties. But the design that worked so well in the Pleistocene is less than optimal in our current environs, where one is much more likely to encounter misinformation than in the environment of evolutionary adaptiveness. Today, as Keith Stanovich (2010) notes, we live in an environment in which other agents may start to arrange the cues to belief in ways that benefit them and not us.

1.4 Obligations

Where does this leave us? In the first section of this paper we argued that the thesis that we can at will acquire beliefs, *of the kind that proponents of doxastic voluntarism have in mind*, is false. The evidence reviewed above show why the italicized qualification is necessary: we can certainly acquire doxastic states that resemble, and may actually qualify as, beliefs, more or less at will. The recipe for acquiring such states is simplicity itself: entertain the proposition that *p* and you will acquire a doxastic state with the content *p*. If states like this count as beliefs, then we can acquire beliefs at will. However, it is apparent that *these* kinds of doxastic states are not the beliefs that philosophers like Frankish think we can acquire in conducive circumstances. They are too unstable and fleeting to be states of the kind that have been at issue in the debate concerning doxastic voluntarism. Frankish suggests as a criterion for beliefs of the latter kind that they serve as premises in *conscious* 'truth-critical deliberation.' States apt to play that role are, inter alia, states that the agent is willing to assert, while the doxastic states acquired automatically are not (always) apt for assertion.

However, these 'thin' states and our propensity to acquire them, when taken together with the knowledge that we have such propensities, do entail that we have obligations, even though doxastic voluntarism—understood in the traditional manner—is false. We can acquire doxastic states of a thin kind at will, and these states affect our behavior.[21] Knowing how we acquire these states imposes obligations. We now know how to acquire thin doxastic states with content *p*: entertain the proposition; the mere fact of having done so will cause you to acquire a doxastic state with a corresponding content, and this state will, in turn, dispose you to come to have what all sides would accept is a

[20] For example, there is evidence that those with a higher 'need-for-cognition' score, do better at rejecting propositions than those with lower scores. See Mandelbaum (2010), particularly the discussion of 'yea-sayers' and 'nay-sayers'.

[21] In particular, they are likely to bias us toward gradually acquiring states with correlative (if not identical) contents that are beliefs on any plausible view.

fully-fledged belief with a matching content. You can increase the likelihood that you will come to believe that *p* by making sure that you are properly distracted so that you don't have time to consider and reject the proposition. The more one encounters the proposition under the requisite load, the more the inferential tentacles the doxastic state will acquire. One can increase their credence in that belief even full well knowing that one is doing that simply by setting up one's environment in a certain way and repeating the above procedure.[22] This set up is unlike Bennett's Credamites set up for we don't need to have anterograde amnesia: we can full well know what we are doing, as long as we are properly distracted. One can have a perfectly well-functioning memory as long as one also has a perfectly well-functioning smart phone to serve as a distracter.

So wherein lies our obligations? We assume here that we can't have obligations over what we cannot control. Since doxastic voluntarism is, strictly speaking, false it appears that we cannot have obligations over what we believe, at least not in any simple or direct way. However, our walk through the empirical literature on belief acquisition pointed to a locus of control we do appear to have over our beliefs: if we are disposed to believe whatever propositions we encounter, then although we may not have direct control over what we believe we do often have control over what ideas we happen to encounter. For example, if we have control over anything, then we have control over what television channel we happen to put on. Suppose you want to watch Fox News because you are interested in seeing how certain types of media portray certain events. Even though this is a benign enough endeavor, you are putting yourself at risk of catching certain beliefs not because the beliefs are worth acquiring epistemically speaking, but rather simply because you encounter them; you run the risk of catching these beliefs in a similar way in which one catches a cold. And just as you can control whether or not you catch a cold to a certain degree—for example, by not kissing someone who has a cold—so too can you control whether or not you encounter, and hence believe, certain propositions.

If the forgoing is correct, we—those of us who know about our propensities to acquire doxastic states through merely entertaining propositions—do have epistemic obligations which arise in the same kind of way in which they would arise were doxastic voluntarism true. We have obligations that arise from the kind of control we actually have over our belief formation process, limited and patchy though it is.

Of course it is often necessary to engage with claims with which one does not agree; even, sometimes, to engage with claims that we know beforehand are dangerously and outrageously wrong. Political scientists, journalists, and cultural critics may all need to watch Fox News for somewhat similar reasons to why physicians expose themselves to infections: for the good of us all. Just as physicians can reduce their risks with proper infection controls, so those who deliberately expose themselves to Fox News

[22] Of course one could accelerate the process by having certain affective variables line up the right way—after all, as the dissonance literature shows we are more inclined to believe what makes us feel better about ourselves (see, e.g., Thibodeau and Aronson 1992).

can take steps to reduce the risk that they acquire the beliefs that Fox disgorges. As we noted above, affirmation leads to belief more reliably when we lack the time and resources to effortfully reject claims. Proper infection control requires that steps be taken to ensure that we are not under cognitive load, stressed, or tired when we enter the quarantine zone.

Unfortunately there are no guarantees that these infection controls will succeed. Given that (a) the affirmation of claims is automatic and (b) affirmed claims *immediately* bias information processing, we can expect even the most fastidious Fox News watcher to acquire attitudes that are influenced by the pollutants they ingest. The viewer may acquire beliefs with the same content as the propositions they encounter or they may instead acquire beliefs that bear the taint of those propositions, either by being entailed by or associated with them.[23] Effortful processing is too slow to keep up with the pace of claim generation.[24] Moreover, though we have picked on Fox as a particularly egregious example of a source of mental contamination, contaminants are ubiquitous. Everywhere in contemporary society there are people attempting to persuade us of claims, to cause us to buy their products or their ideas. We encounter many of these messages when we are under load: stressed, tired, or distracted (when we are commuting and quickly pass by a billboard, for instance).[25]

Further, even if the agent is able to avoid mental contamination near the time of exposure, the danger has not passed. Even after we have evaluated claims, with whatever degree of success, we remain vulnerable to psychological mechanisms that leave us with unjustified beliefs. The 'sleeper effect' produces a delayed increase in the persuasiveness of a claim (Pratkanis et al. 1988). Sometimes a message is presented together with a discounting cue—e.g., message: 'global warming is a myth'; discounting cue: source of message is the oil industry. Subjects who evaluate the message may initially give it little weight, but their confidence in its truth tends to rise over time. Why? One possibility is that when the message is recalled, the discounting cue is not, because there is only a weak association between message and cue; the message and the cue may be stored in different memory networks such that their inferential connections, and thus their decay rate, will differ.[26] Since messages are often much more

[23] In fact, the situation where one acquires a belief that is related but not identical to the content of the original perception may be more dangerous than one where the original content is believed straight away, for the former situation increases the likelihood of the subject failing to recall the source of the belief (thus increasing the 'sleeper effect'; see later in this section).

[24] It is also worth noting that some of the techniques Fox and other news organizations employ, such as the use of simultaneous scrolling text and unrelated news delivered verbally, could not be better designed to induce cognitive load; indeed they closely resemble standard methods used in social psychology to this end (see Mandelbaum (2010) for the gory details). This fact will make efforts at good cognitive hygiene all the more likely to fail.

[25] It is important to note how light the required load can be. Merely self-regulating one's own behavior is often enough load to be distracting and accelerate the quick acquisition effect (Gilbert 2002). Of course, in social situations (including the classroom!) one is, often enough, trying to self-regulate for fairly mundane reasons.

[26] A similar style of explanation can be used to explain the 'source monitoring' errors underwriting the (false) recovered memory phenomenon (Schacter et al. 1997). Source monitoring (i.e., recalling the source

vivid, they tend to be more accessible and available for recall. Consequently memory for the discounting cue can decay more quickly than memory for the claim, since the discounting cue is apt to not be as integrated and activated as the message is (Kumkale and Albarracín 2004).

1.5 Conclusion

Though doxastic voluntarism is false, we nevertheless have epistemic obligations generated in much the same kind of way in which its truth would generate epistemic obligations. Were doxastic voluntarism true—i.e., were we to have the kind of direct control over the content of our belief that it entails—then we would have obligations to use this power well, and might reasonably be praised or blamed for (some of) the beliefs we formed. We do not have direct and immediate control over the content of our beliefs in the way envisaged, but we do have some degree of control: we make it likely that we will acquire beliefs by mere exposure to them. Just as we have obligations to take risks into account when we act, we have obligations to take the risk of forming unjustified and, worse, immoral beliefs into account when we expose ourselves to them.

We can do various things to reduce the risks, but we cannot reduce them to zero except by avoiding exposure altogether. Of course when we decide how to act we need to weigh the risks against benefits: that an action carries with it a potential risk of harm does not entail that the action is impermissible, or even inadvisable. Everything depends on the magnitude of the harms, the probabilities of avoiding them, and the

of a signal) is particularly important in cases of recovered memories of abuse. In these cases, a therapist cues patients and prods them to remember (or 'remember') traumatic experiences that they have forgotten (or 'forgotten'). Although it's unclear whether any of these cases of recovered traumatic memory are veridical, it is clear that many of the supposed cases of recovered traumatic memory are not veridical. In these cases, the patients create, rather than recall, the event. The patient comes to 'recall' the event only after a therapist's suggestion; because they fail to appropriately monitor the source of the memory, they take suggestion for recall (the Gilbert style experiments can be interpreted as presenting cases where subjects forget the source tags of 'true' and 'false'). With regard to recovered 'memories', the effects may be potentiated by other features of the context. For instance, load can be brought about by the mere intensity of the situation (being asked to recall traumatic events). The problem then metastasizes because of the stereotypes normally invoked in this recall and the involvement of episodic memory. The patient generally has some negative feelings built up generally towards an older male figure, like a father, uncle, or priest. These figures have quite stereotypical traits that are easily conjured up. The combination of stereotype activation and cognitive load make for a volatile situation. In a study on stereotypes and source monitoring Sherman and Bessenoff (1999) found that when under cognitive load, participants are apt to default to judgments that fit a stereotype even if they were just shown that the stereotype does not hold for the case at hand. The interaction between stereotypes and cognitive load in recovered memories situations is exacerbated because episodic recollection is more demanding and effortful than semantic recollection (Tulving 1983). When patients are asked to recall traumatic memories, they are being asked to recall episodic memories and are thus put under additional load, making faithful memory search quite difficult. Semantic recollection, on the other hand, is much less effortful and can occur under load. Thus, when people are put under load they are apt to resort to the stereotypes that are stored in semantic memory while lacking access to their actual episodic memories.

magnitude and probabilities of potential benefits. We think that there is an obligation on us to take all these facts into account when we act.

References

Adams, R. M. (1985). 'Involuntary Sins', *Philosophical Review* 94: 3–31.

Alston, W. (1996). *A Realist Conception of Truth*. Ithaca, N.Y.: Cornell University Press.

Bennett, J. (1990). 'Why Is Belief Involuntary?' *Analysis* 50: 87–107.

Bratman, M. (1992). 'Shared Cooperative Activity', *The Philosophical Review* 101: 327–41.

Brock, T., and Balloun, J. (1967). 'Behavioral Receptivity to Dissonant Information', *Journal of Personality and Social Psychology* 6(4): 413–28.

Feldman, R. (2001). 'Voluntary Belief and Epistemic Evaluation', in M. Steup (Ed.), *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. New York: Oxford University Press, 77–92.

Frankish, K. (2007). 'Deciding to Believe Again', *Mind* 116: 523–47.

Gilbert, D., Krull, D., and Malone, M. (1990). 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information', *Journal of Personality and Social Psychology* 59(4): 601–13.

Gilbert, D. (2002). 'Inferential Correction', in T. Gilovich, D. Griffin, and D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press, 167–84.

Gilbert, D., Tafarodi, R. and Malone, P. (1993). 'You Can't Not Believe Everything You Read', *Journal of Personality and Social Psychology* 65 (2): 221–33.

Ginet, C. (2001). 'Deciding to Believe', in M. Steup (Ed.), *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. New York: Oxford University Press, 63–76.

Hastorf A., and Cantril, H. (1954). 'They Saw a Game: A Case Study', *The Journal of Abnormal and Social Psychology* 49(1): 129–34.

Hieronymi, P. (2006). 'Controlling Attitudes', *Pacific Philosophical Quarterly* 87(1): 45–74.

Holton, R. (forthcoming). 'Intention as a Model for Belief', in M. Vargas and G. Yaffe (Eds.), *Rational and Social Agency: Essays on the Philosophy of Michael Bratman*. Oxford: Oxford University Press.

Klayman, J., and Ha, Y. (1987). 'Confirmation, Disconfirmation, and Information in Hypothesis Testing', *Psychological Review* 94(2): 211–28.

Kumkale, G. T., and Albarracín, D. (2004). 'The Sleeper Effect in Persuasion: A Meta- Analytic Review', *Psychological Bulletin* 130(1): 143–72.

Levy, N. (2005). 'The Good, the Bad and the Blameworthy', *Journal of Ethics and Social Philosophy* 1:1–16.

Levy, N. (2011). *Hard Luck*. Oxford: Oxford University Press.

Mandelbaum, E. (2010). 'Thinking is Believing: An Essay on the Unbearable Automaticity of Believing', Ph.D. Dissertation, University of North Carolina, Chapel Hill.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., et al. (2010). 'Willful Modulation of Brain Activity in Disorders of Consciousness', *New England Journal of Medicine* 362: 579–89.

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., and Pickard, J. D. (2006). 'Detecting Awareness in the Vegetative State', *Science* 313: 1402.

AQ: Please provide page range for the reference - Forthcoming: OUP online indicates June 2014 (estimated)

Pratkanis, A.R., Greenwald, A.G., Leippe, M.R., and Baumgardner, M.H. (1988). 'In search of reliable persuasion effects: III. The sleeper effect is dead. Long live the sleeper effect', *Journal of Personality and Social Psychology* 54(2): 203–18.

Ross, L., Lepper, M., and Hubbard, M. (1975). 'Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm', *Journal of Personality and Social Psychology* 32 (5): 880–92.

Schacter, D.L., Norman, K.A., and Koutstaal, W. (1997). 'The recovered memories debate: A cognitive neuroscience perspective', In M. Conway (Ed.), *False and recovered memories*. New York: Oxford University Press.

Schwitzgebel, E. (2002). 'A Phenomenal, Dispositional Account of Belief'. *Nous* 36: 249–75.

Setiya, K. (2008). 'Believing At Will', *Midwest Studies in Philosophy* 32(1): 36–52.

Sherman, J., and Bessenoff, G. (1999). 'Stereotypes as Source-Monitoring Cues: On the Interaction between Episodic and Semantic Memory', *Psychological Science* 10 (2): 106–10

Smith, A. M. (2005). 'Responsibility for Attitudes: Activity and Passivity in Mental Life', *Ethics* 115: 236–71.

Smith, H. (1983). 'Culpable Ignorance', *Philosophical Review* 92: 543–71.

Stanovich, K. (2010). *Rationality and the Reflective Mind*. Oxford: Oxford University Press.

Steup, M. (2000). 'Doxastic Voluntarism and Epistemic Deontology', *Acta Analytica* 15: 25–56.

Stich, S. (1978). 'Beliefs and Subdoxastic States', *Philosophy of Science* 45: 499–518.

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

Tulving, E. (1983). *Elements of Episodic Memory*. New York: Oxford University Press.

Wegner, D., Coulton, G., and Wenzloff, R. (1985). 'The Transparency of Denial: Briefing in the Debriefing Paradigm', *Journal of Personality and Social Psychology* 49 (2): 338–46.