A Defense of Back-End Doxastic Voluntarism

Laura K. Soter

Penultimate draft. Please cite published version: http://doi.org/10.1111/nous.12501

Doxastic involuntarism—the thesis that we lack direct voluntary (non-evidential) control over our belief states—is often touted as philosophical orthodoxy. I here offer a novel defense of doxastic voluntarism, centered around three key moves. First, I point out that belief has two core functional roles, but that discussions of voluntarism have largely ignored questions of control over belief's guidance function. Second, I propose that we can learn much about doxastic control by looking to cognitive scientific research on control over other relevantly similar mental states. I introduce a mechanistic account of guidance-control for "emotion-type states," and argue that these same cognitive control mechanisms can used to control doxastic guidance—what I call "back-end" voluntarism. Crucially, these back-end control mechanisms can be deployed in response to reasons which are not the right kinds of reasons to support "front-end" belief formation—that is, back-end control is deployable for non-evidential reasons. Third, I argue that comprehensive, self-directed exercises of this kind of control can amount to an underappreciated kind of voluntarism. I discuss upshots of the view, including aspects of the psychological profile it brings to light, and implications for various philosophical debates.

1. Introduction

There is a longstanding debate around whether, and in what sense, we have voluntary control over our belief states. *Doxastic voluntarism* is the thesis that our beliefs are under our voluntary control—that we can, in some sense, believe at will. Doxastic *in*voluntarism is the denial of this thesis. Involuntarism is often heralded as the orthodox view: most philosophers hold that we cannot simply choose what we believe; rather, beliefs are thought to be (psychologically, conceptually, or normatively) constrained by our evidence.

In this paper, I offer a novel defense of doxastic voluntarism. Although standard involuntarist perspectives capture something important about the nature of belief, these arguments do not extinguish the prospects for voluntarism. I highlight that the in/voluntarism debate has focused nearly exclusively on "front-end" control over belief's *evidence-responsiveness* function; this focus has obscured the question of whether we have "back-end" control over belief's *guidance* function. I offer a defense of, and an account of the mechanisms involved in, back-end control, drawing on control mechanisms that are well-studied by cognitive scientists but have been overlooked in the context of belief, and arguing that thoroughgoing back-end control can amount to an exercise of direct doxastic voluntarism.

In doing so, I am to offer an account of doxastic control that is empirically plausible and mechanistically specific, and which seeks to bring discussions of voluntarism out of the armchair and into contact with the cognitive science of mental control.

This paper proceeds as follows. First, I describe the contours of the existing voluntarism debate (§2), and highlight a common view about the two key functional roles of belief (§3). I point out that the debate has focused on the question of control over just one of these functional roles, eliding consideration of the other (§4), and argue that this overlooks another possible route to voluntarism. I then propose that a fruitful way to tackle the question of doxastic control is to consider what we know about control over other kinds of relevantly similar mental states. I consider a class of mental states that share a functional architecture with belief, and highlight a recent analysis of the kinds of control we do and don't have over these states (§5). I then return to beliefs, using the prior discussion to develop an account of "back-end" doxastic voluntarism (§6). I close with some objections to (§7) and upshots of (§8) the proposal.

2. Involuntarism as Orthodoxy

Let's begin by laying out the state of the traditional doxastic voluntarism debate. The key issue can be specified in various ways, but it circles around questions such as: whether we have voluntary control over our beliefs, whether we can choose what to believe, or whether we can believe at will. Broadly, **voluntarism** gives affirmative answers to these questions, while **involuntarism** gives negative ones: voluntarism holds that we do have voluntary control of our beliefs, and involuntarism holds that we do not.

Involuntarism is generally accepted as the orthodox view.[1] More precisely, the standard view is that we cannot choose to believe in response to just any kind of reason that we take to weigh in favor of having a belief;[2] instead, beliefs are (rationally and/or psychologically) constrained by *evidence*—by information that (an agent takes to) bear on the truth or falsity of a proposition. Many have motivated this constraint by appealing to felt psychological limitations (Alston, 1988; Chrisman,

[1] See Jackson (2021), Levy & Mandelbaum (2014), Roeber (2019, 2020).
[2] Hieronymi (2006) calls these "extrinsic" reasons: reasons that show a belief would be good to have.

2008, 2022; Williams, 1973): no matter how beneficial it would be for us to hold some belief, if that belief is not supported by our evidence, we simply seem to lack the psychological capacity to form that belief. This is the "No Rewards Principle" (Chrisman, 2008): if you offered me one million dollars to believe that San Francisco is in Minnesota, despite it being to my benefit to believe this, I do not seem able to believe in response to that (compelling) practical, non-evidential reason.

The explanation for this is often given in terms of the nature of belief. Belief, in some sense, "aims at truth:"[3] it represents its object as true, is formed in response to evidence that bears on the truth or falsity of the claim in question, and it is regulated (and taken by agents themselves to be governed) in response to of truth as a standard of correctness (Shah & Velleman, 2005; Velleman, 2014). Some describe this idea as belief being "commitment-constituted" (Hieronymi, 2006; Singh, forthcoming): believing that *p* involves a commitment to *p*'s truth that leaves you answerable to these evidentialist rational standards. This deep tie between truth and belief motivates involuntarism: an agent cannot, it seems, seriously regard some mental state of hers as a *belief* if she knows that it was formed on the basis of reasons that have nothing to do with its truth and she takes that state to be unsupported by evidence (famously, Williams, 1973; see also Frankish, 2007; Levy & Mandelbaum, 2014; Scott-Kakures, 1994[4]). Though literatures are dedicated to fine-tuning each of these points, the basic ideas have widespread support; on their basis, many philosophers have agreed that we cannot choose to believe in response to practical and/or moral considerations which have no bearing on the truth of the proposition at hand.

There are two critical qualifications to the claim that involuntarism is the standard view. First, the arguments described aim to show that we cannot believe in response to any kind of reason we want, with no regards for the truth of the matter in question. There is, however, a notable camp of philosophers who defend a self-labeled *voluntarist* thesis that takes a broadly compatibilist approach to

[3] Classically, See Williams (1973) and Velleman (2000). There is controversy over precisely how to understand this idea; see Shah (2003), Shah and Velleman (2005), and Wedgwood (2002) for challenges and Singh (forthcoming) for an attempt to reconcile it.

[4] Frankish (2007) refines these ideas from Williams. Scott-Kakures (1994) describes similar ideas as the *self-defeatingness gambit*. See Bennett (1990) for criticism; and Singh's (forthcoming) note that agents could misapply the concept of belief.

doxastic freedom, arguing that that the will is efficacious in our believing even though we cannot believe "willy nilly" (the technical term for believing however we wish). Belief is nonetheless voluntary or agential, these authors argue, because when we believe in response to our evidence, we believe *as we mean to*—we believe in response to the right kinds of reasons for belief, namely those that help settle the matter of whether *p* (see Ryan (2003), Steup (2012, 2017, 2018), Hieronymi (2006, 2008), and Singh (forthcoming) for versions of this view. Note that some of these authors disagree in their labels in part because of disputes over what counts as "voluntary;" e.g., Hieronymi denies that this amounts to *voluntary* control, but agrees we have *evaluative* control.[5]). Notably, endorsing this flavor of voluntarism does not depend on thinking that we can choose to believe in response to non-evidential reasons; these two kinds of voluntarist thesis are conceptually distinct.[6]

Second, the standard arguments aim to show that we lack *direct* voluntary control over our beliefs: we cannot immediately and directly believe in response to non-evidential reasons. In contrast, everyone agrees that we have various kinds of *indirect* doxastic control.[7] We can exercise indirect control when we have the power to change the truth value of the proposition in question (if I want to believe that the lights are on, I just have to flip the light switch; Feldman, 2000). More substantively, we can exert a great deal of control over the processes of inquiry, reasoning, and evidence-gathering that lead to what we believe: we can choose whether to inquire into some topic, how much and what kind of evidence to gather, how carefully and exhaustively to think through a question, etc.,[8] and we can do all of this for practical reasons.

We might model this distinction visually, in Figure 1:

---

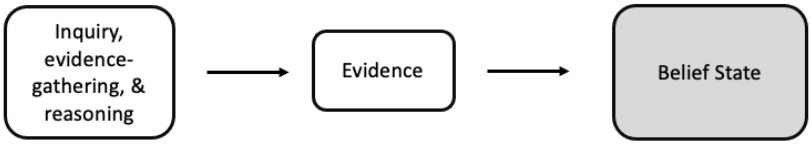
[5] Hieronymi (2006; 2008) argues that belief is not voluntary because it does not result from forming an intention, and cannot respond to any reasons we take to show believing worth doing, which she takes as a constraint on voluntary action. In contrast, Ryan, Steup, and Singh defend hold that belief's reasons-responsiveness is still voluntary, on a certain way of understanding that notion, as does Shah (2002).

[6] They are sometimes conflated, likely in part because Alston (1988) uses observations about our lack of non-evidential voluntarism to draw skeptical conclusions about our lack of broader doxastic agency. See Singh (forthcoming) for a nice analysis of the problems with this move.

[7] Direct control is often understood in terms of basic action, where forming the belief is undertaken immediately in response to the relevant (non-evidential) reasons and without having to do anything else (Levy & Mandelbaum, 2014). There is contention over precisely how to understand the direct/indirect distinction; see Jackson (2021), Frankish (2007), and Hieronymi (2009) for discussion, and Vermaire (2022) for skepticism.

[8] This is sometimes labeled "belief management" (Floweree 2020; Hieronymi, 2006).

Figure 1. *(The Beginnings of) A Process Model of Belief*



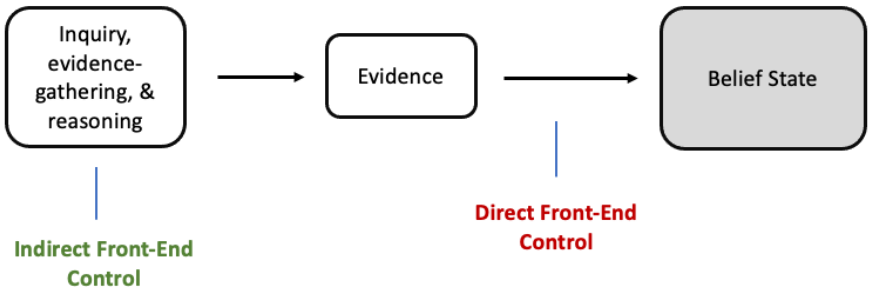
Here, "inquiry" refers broadly to practices of evidence-gathering, reasoning, and thinking through a matter: the "upstream" practices that lead us to have the body of evidence bearing on *p* which we come to assess; "belief state" is the representational state, our confidence in the truth or falsity of *p*, that arises from that evidence-assessment process. Note that here and throughout, the belief state in question can be any level of epistemic confidence in *p*; the preceding arguments regarding in/voluntarism do not rely a strict notion of "full" or "on/off" belief.

So: the traditional view acknowledges that we can exert plenty of control over the inquiry process: this is *indirect* doxastic voluntarism, through which we alter our beliefs by intervening on what evidence we have or how we think about it.[9] But once we hold fixed the evidence and interpretation of it, that leads directly, ballistically, and automatically to the belief state in question. We cannot intervene here in response to non-evidential reasons, because, on the traditional view, a belief state *just is* a reflection of our assessment of the evidence: there is something like a psychological entailment relationship between the appraisal of the evidence and the belief state (see Arpaly, forthcoming). We can represent these standard views in Figure 2:[10]

[9] The scope of our indirect control is arguably limited; it's not clear that indirect control can reliably bring about a *precise* desired belief state (Hieronymi, 2006, p. 55); intervening with such a goal risks being self-undermining (Williams, 1973).

[10] This model also allows us to locate some recent challenges to traditional involuntarism by those who hold that belief is not just a reflection of evidence, but that plus something else. Some have emphasized the act of judgment, and argued that (in epistemically permissive cases where the evidence underdetermines the rational doxastic state)—an agent can decide whether to (for instance) believe *p* or withhold judgment about *p* (Frankish, 2007; Jackson, 2021; Roeber, 2019, 2020; also Quanbeck & Worsnip, forthcoming; though see Kieval (2022) and Sylvan (2016) for critical discussion). Others think believing involves active decisions such as closing inquiry (e.g., Friedman, 2019). Some propose these views offer new routes to (direct) voluntarism, because we can make such choices (not to close inquiry, or to withhold judgment instead of believing) for practical reasons. I set these views aside for present purposes; how the debate about traditional direct voluntarism ultimately shakes out does not affect my positive argument.

Figure 2. *The Standard View on Voluntarism: Endorsing Indirect Voluntarism, Denying Indirect Voluntarism*



Note. The colors represent the orthodox views: indirect voluntarism is in green, as everyone agrees we have voluntary control there; direct voluntarism is in red, as most deny that we can voluntarily intervene there.

This provides us with a good gloss on the status quo. Granting some key nuances, the orthodox view is that, because belief states arise automatically and directly in response to evidence, we cannot change those states directly in response to non-truth-relevant reasons. Thus, we lack direct voluntary control over our belief states, and involuntarism reigns.

I will argue that this move—from the claim that belief states respond ballistically and specifically in response to evidence, to the involuntarist conclusion—is too quick. To see why, we need to say a bit more about the nature of beliefs and how they function.

3. Two Functions for Belief

Belief is often thought to have two central functional roles in our cognitive economies. Call the first **appraisal**:[11] beliefs form and change in response to evidence. This idea can be fleshed out both psychologically and rationally. Philosophers often describe beliefs as constitutively evidence-responsive: (part of) what it is for a mental state to count as a belief state is for it to be the kind of state that reflects our evidence and changes (spontaneously and non-inferentially) when that evidence

[11] I use this term for the sake of continuity with the psychological literature on emotions; the relevance of this will become apparent shortly.

changes.[12] This function is tightly related to the characterization of belief as subject to epistemic standards of correctness and truth: beliefs are successful, accurate, or rational to the extent that they reflect our evidence (Railton, 2014; Shah & Velleman, 2005; Wedgwood, 2002). This idea underpins "evidentialist" norms on rational belief (the normative view that only evidential reasons are rational reasons for belief; e.g., Kelly, 2002; Shah, 2006).

Call the second function **guidance:** beliefs serve as our "default cognitive background" (Bratman, 1992)—they (spontaneously and non-inferentially) guide various psychological and behavioral processes, including our patterns of planning, deliberation, inference, action, reasoning, and so on (see Schwitzgebel, 2006, Railton 2014). That is, once we have appraisal-generated representations, those confidence states do not sit inert in our psychologies; rather, they have a wide range of cognitive and behavioral effects. Our appraisals of the evidence (in interaction with other mental states) affect our patterns of attention, memory, thought, motivation, judgment and inference, goal selection, action tendencies, etc., often automatically and without our direct oversight. Beliefs are, in other words, prepotent mental states that cause automatic effects across a diverse range of psychological mechanisms.

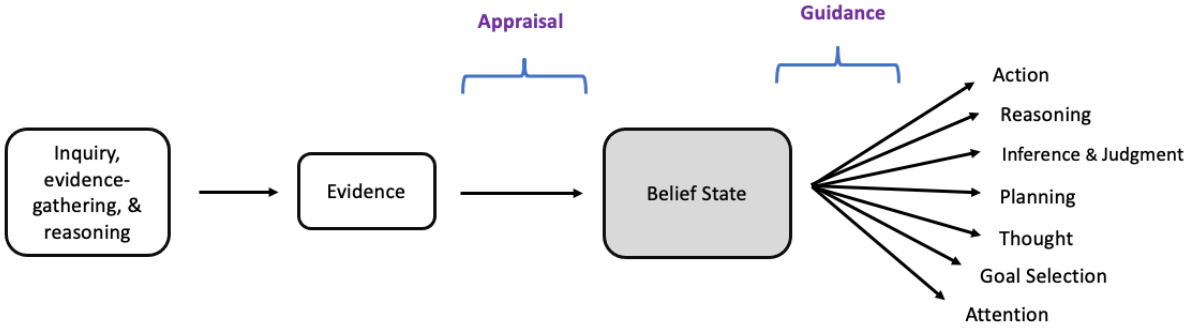
The centrality of both appraisal and guidance is widely recognized by various philosophical theories of belief (though not necessarily using these labels). There is a strong tradition of thinking that the proper functioning of both components is central to or constitutive of believing. Versions of this appraisal-guidance structure can be found in various traditions, including classic metasemantic theories that characterize beliefs in terms of both the information they capture and reflect and the way they determine outputs in behavior and cognition (Dretske, 1991; Lewis, 1974), contemporary representationalist theories that appeal to both the ways in which belief-representations respond to certain kinds of inputs and use those representations in various downstream computations and psychological processes (Porot & Mandelbaum, 2021; Quilty-Dunn & Mandelbaum, 2018), and

[12] For recent defenses, see Flores (forthcoming), Helton (2020), and Singh (forthcoming); also Egan (2008), Shah and Velleman (2005), Levy (2015), Shah (2003), Velleman (2000), Davison (1985), Railton (2014), and Wedgwood (2002, 2007).

epistemological normativist theories that claim we need both "input side" and "output side" constitutive norms of belief (Nolfi, 2015). Often (though not always) appreciation of the appraisal and guidance roles is thought of in functionalist terms. Functionalism holds that what it is for something to be a mental state of a particular kind is for it to play the characteristic functional role(s) of that state in a cognitive system (Block, 1980; Levin, 2021). In the context of belief, the functionalist thought is that what it is for something to be a belief state is for it to be the kind of thing that typically or characteristically arises in response to evidence, and typically or characteristically shapes reasoning, cognition, and action in relevant ways (see Schwitzgebel, 2011). Beyond direct endorsements of the importance of both roles, we can also appreciate the widespread influence of this architecture indirectly. One piece of evidence comes from noticing that whenever philosophers encounter an odd mental state that fulfills one role but not the other, such as implicit biases or delusions,[13] there is reliably an accompanying debate about whether these are really *beliefs*.

Across these views, there is a shared thought: for something to be fully characterized as a belief, it must play both the appraisal and guidance roles. We can expand our model of belief to include guidance, in Figure 3:

Figure 3. *The Two-Pronged Architecture of Belief*



Note. The processes stemming from guidance are meant to be representative but not exhaustive or precisely taxonomized, especially in their relations to each other (e.g., action often comes after, and as

[13] For discussion of implicit bias, see Gendler (2008b; 2008a), Levy (2015), Mandelbaum (2016), and Madva (2016). For discussion of delusion, see Bortolotti (2005), Bortolotti and Miyazono (2015), Egan (2008), and Flores (2021).

a result of, the other processes, which themselves can be broken into more precise mechanistic components and processes).

Although the two-pronged view is widely held among philosophers, it is not universally adopted; some think that belief should be understood just or primarily in virtue of only one component. Some think appraisal is what really matters for belief; this is arguably (though perhaps implicitly) a dominant view in some areas of epistemology, where it's common to see characterizations of belief states *just as* states that reflect an agent's confidence in some proposition.[14] Given the field's focus on rational standards of belief-formation and evidence-responsiveness, what happens after appraisal is sometimes seen as a secondary question to what the agent believes.[15] In contrast, others—notably dispositionalist and some pragmatist theories of belief—think that *guidance* is what really characterizes belief. Broadly, these theories hold what an agent believes is defined in terms of how she is disposed to think, reason, and act in various contexts (e.g., Peirce, 1878; Ryle, 1949; Schwitzgebel, 2002; Zimmerman, 2018, *inter alia*); guidance is the crucial function for characterizing an agent's beliefs. Though guidance-focused views have received slightly less attention in epistemology, dispositionalism in particular is quite prominent in philosophy of mind. Dispositionalism can be characterized as a "one-pronged" functionalist view, insofar it holds that guidance is the only function that matters for determining what an agent believes (Schwitzgebel 2011).

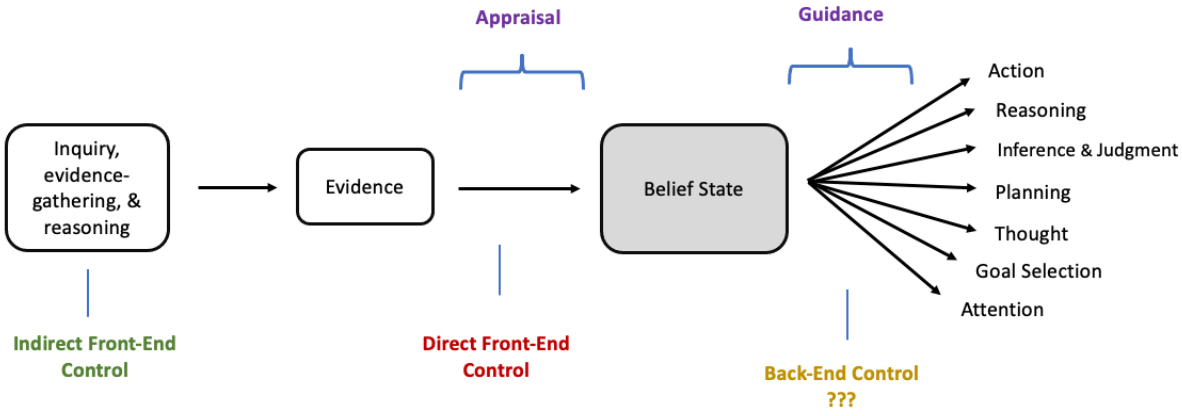
4. Another Route to Voluntarism?

We can now notice something striking about the voluntarism literature: the classic debate focuses nearly exclusively on control over the appraisal component of belief. The question of control over belief's guidance component has been nearly entirely overlooked. But we can now appreciate that there are really two possible prongs for questions of doxastic control: we can ask whether we have

[14] E.g., Wedgwood's (2002, 2007) normativism puts constitutive norms only on the "input side."
[15] Indeed, some epistemologists think that once we start to think about guidance, we're not even really talking about belief anymore, but instead some other attitude like acceptance: see, e.g., Cohen (1989, 1995) and Begby (2021, Chs. 1 and 9)—on these views, acceptance is the attitude that guides reasoning and action.

front-end control over the evidence-responsive appraisal mechanism, or we can ask whether we have **back-end control** over whether and how beliefs guide various processes across reasoning, cognition, and action. We can thus update our map of the debate:

Figure 4. *Different Questions of Control*



The debate over front-end control is well-trodden: as discussed in §2, we have various kinds of indirect front-end control, but (many hold) we lack direct front-end control. But the question of back-end doxastic control—whether we can exert systematic control over the guidance function, and what mechanisms might be involved in so doing—has not been seriously considered.

This is striking, given that we just highlighted that on classic and influential views, guidance is a (or even *the*) constitutive role of belief: (at least) part of what it is for an agent to believe *p* is for the *p*-appraisal to guide various processes across reasoning, cognition, and action. This opens space for a different kind of voluntarism thesis: if we can systematically prevent a state from instantiating its guidance function, we can thereby prevent the state from fulfilling a constitutive function of belief—thus challenging its status *as* a belief. So even if arguments about the lack of direct front-end control block one possible route to voluntarism, they do not yet justify a full involuntarist conclusion: there remains room for the possibility of direct voluntary (non-evidential) control on the back-end.

Making good on this possibility requires two things: first, telling a psychologically plausible story about the mechanisms of back-end doxastic control; and second, showing that back-end control can indeed amount to an exercise of voluntarism. I'll take on these two tasks in what remains.

5. Relevantly Similar Mental States: Two-Pronged Architecture beyond Belief

The first thing we need is an account of what back-end doxastic control might look like, cognitively. To this end, I make a methodological proposal: we can learn about doxastic control by looking to what we know about the contours (and limits) of control over other relevantly similar mental states. The operative sense of similarity here will be in sharing the "two-pronged" appraisal/guidance architecture. If we look beyond beliefs, we'll quickly realize this is a familiar functional architecture shared by many other mental states—and that we know quite a lot from cognitive science about what kinds of control we do (and don't) have over such states. If it turns out that existing cognitive scientific frameworks offer resources that might be applicable to questions of doxastic control, this provides an opportunity to develop an account that is independently psychologically plausible, insofar as it appeals to scientifically supported cognitive architectures and mechanisms. Let's thus take a brief detour away from beliefs.

5.1. The Architecture of Emotion-Type States

One useful recent account comes Sripada (2021)'s discussion of "emotion-type states." These states are unified, on his view, precisely in virtue of their shared functional architecture. The group includes emotions as paradigmatic instances, along with other states like drives, impulses, and cravings, among others. These states have an **appraisal** function**:** they are elicited spontaneously and automatically in response to (perceived) state-relevant stimuli (Ellsworth, 2013; Ellsworth & Scherer, 2003; Moors, 2014; Roseman & Smith, 2001).[16] If an agent encounters a situation appraised as threatening, fear is elicited; if she encounters a stimulus perceived as unclean or contaminated, disgust is elicited; and so on. Once elicited, these states cause widespread downstream default cognitive, physiological, and behavioral consequences (Adolphs & Andler, 2018; Keltner & Gross, 1999)—that is, they have a **guidance** function**.**[17]

[16] That these appraisals are spontaneous and nondeliberative does not mean that they are "brute," "nonrational," or otherwise non-responsive to reasons. Rather, emotions are selectively elicited in a (at least quasi-) rational relationship to the stimuli that make sense given those emotions (see Roseman & Smith, 2001; D'Arms & Jacobson, 2000, 2023).

[17] Sripada (2021)'s architecture is displayed in a figure (p.809) that readers will notice the figures in this paper parallel.

Sripada discusses the nature of the guidance function in detail; I'll here give a gloss highlighting key features. Activated emotion-type states lead to *default state-congruent biases or effects across a wide range of psychological mechanisms*: emotions affect our default patterns of attention (Phelps et al., 2006), how we evaluate situations and information, the inferences we draw and the beliefs we form (Angie et al., 2011; Bodenhausen et al., 1994; Bower, 1991; Lerner & Keltner, 2000); what we recall and encode into memory (Hamann, 2001; Kensinger, 2009); goal selection and action motivation (Frijda, 1986, 1987; Scarantino, 2014); what thoughts spontaneously come to mind (Bower & Cohen, 1982; Smallwood et al., 2009), and so on. If fear is activated for an agent, her default patterns of cognition and action will be guided in predictable ways: she'll be biased, e.g., towards escape-related goal selection (Frijda, 1987; Frijda et al., 1989), be faster and more likely to attend to threat-relevant stimuli (Öhman et al., 2001), spontaneously evaluate ambiguous stimuli as threatening (Eysenck et al., 1991) be likely to call to mind threat-relevant thoughts and memories (Smallwood et al., 2009), etc.

Sripada characterizes these biases as instantiated via the activation of a series of *response pulses*: brief, simple states that are impulses towards a particular response by a psychological mechanism in a specific stimulus context. When a response pulse is activated, the associated response will occur by default, unless some exogenous force intervenes. For instance, under many conditions people experience a response pulse to shift their gaze towards a moving object in a still scene: detecting movement activates an impulse to look towards the movement, and this gaze shift occurs by default, unless the agent intervenes to prevent it. Activated emotion-type states give rise to emotion-congruent response pulses across the various cognitive mechanisms discussed above; and these extended streams of biased response pulses across mechanisms and over time build patterns of emotion-type state congruent reasoning, cognition, and action. In other words, they instantiate the emotion-type state's guidance function.

5.2. Control and Emotion-Type States

Sripada lays out this architecture in service of explicating control over emotion-type states. There are some things we cannot control. We cannot control how the appraisal process gives rise to the state (at time *t* with appraisal conditions *a*, we cannot prevent *a* from giving rise to emotion-type

state *e*); this process is automatic and ballistic. We thus cannot directly will an emotion-type state in or out of existence independently of the appraisal processes for state-irrelevant reasons; if we appraise a stimulus as frightening, we cannot simply and directly will away the fear, or will ourselves to find it funny.[18] Additionally, we cannot control that the appraisal will *by default* elicit state-congruent responses: i.e., we cannot directly prevent the response pulses from being activated.

But there are places where we *can* exert control. In particular, we can control whether the default responses are *actualized*: we can exert **back-end interventionist control**, preventing the associated responses from realization. Perhaps I'm afraid of a spider I've seen scuttle across my floor, and biases me towards (among other things) shifting my gaze to the dusty corners where critters tend to lurk. That will be my default fear-guided responses—but clearly I don't *have* to shift my gaze; I can override that default.

We block such responses by deploying **cognitive control mechanisms:** basic mental control acts that override and redirect default psychological responses. These mechanisms are often studied via "conflict tasks:" experimental setups that produce a characteristic divergence between an automatic, spontaneous response to a stimulus, and the response demanded by task instructions. A classic example is the Stroop Task (Stroop, 1935), in which participants see color words ("green") printed in colored text, and are instructed to state the text color of the word. On congruent trials, the text color and the word are the same ("green" in green text); on incongruent trials, they differ ("green" in red text). (Due to a deeply learned habit of word-reading in literate adults), there is a spontaneous tendency (response pulse) to read the word; on incongruent trials, the participant must override the spontaneous word-reading tendency to follow the task instructions and report the text color.[19] This

[18] This is not to say it is impossible for us to intervene on the "front-end" of emotion; e.g., "reappraisal" strategies involve reframing a stimulus such that it changes (in kind or strength) the elicited emotion. On our process model, this is ultimately a kind of (proximate but) indirect front-end control: the agent works to change how the stimulus is represented, such that a different emotion is elicited; they still do not intervene directly on the appraisal-to-elicitation process. I ultimately think this is also precisely what we do when we "think differently" or "reframe the evidence" to elicit a different belief state, which philosophers often loosely talk about. In another working paper, I tackle applying this broader framework to doxastic control; here I remain focused on the back-end, because this is the neglected issue. Moreover, reappraisal also involves the same cognitive control mechanisms I will discuss shortly, so this is a useful starting point.

[19] Other well-known conflict tasks include anti-saccade tasks, flanker tasks, go/go tasks, and think/no think tasks, among many others. Though at a high level, the operations involved in each of these tasks look quite different—involving

overriding, inhibition, and redirection of the default response is an exercise of cognitive control. Cognitive control is characteristically effortful, engaging executive processes to override spontaneous responses and redirect them towards goal-congruent ones.

So: emotion-type states produce state-congruent biases in the form of response pulses across a range of cognitive mechanisms. We can prevent response pulses from actualizing their associated responses via the deployment of effortful cognitive control mechanisms. Putting these together: we deploy back-end control over emotion-type states, by executing skilled sequences of cognitive control actions, to prevent them from instantiating their guidance function across psychological and behavioral mechanisms. Crucially, we do this when some appraisal-elicited state conflicts with our goals or practical motivations. Though appraisal is constrained by state-relevant reasons/stimuli, the deployment of back-end control to suppress these states is *not* so constrained. I can't directly will away a spider-fear-appraisal because I want to look cool in front of a crush, but I *can* exert systematic back-end control to block fear-guidance because I want to look cool.

This section provided a lot of technical apparatus rather quickly. Really, we just need two takeaways: (1) emotion-type states have an appraisal and a guidance function, and the guidance role is instantiated via default state-congruent effects across a diverse range of cognitive mechanisms. (2) We can intervene on the guidance role, when so motivated, via the (skilled, effortful) deployment of cognitive control mechanisms. One final piece of Sripada's framework is that these mechanisms constitute exercises of **self-control**: to exert self-control is, on his view, to deploy extended streams of cognitive control acts to prevent these state-congruent default cognitive and behavioral effects, redirecting those mechanisms towards responses more consistent with our presently held goals or commitments.

With that, let's return to beliefs.

---

mechanisms of visual attention, mental association, motor response, word-reading—they each capitalize on the same mechanistic structure: a response pulse must be effortfully overridden when task-appropriate.

6. Back to Beliefs: Back-End Doxastic Control

Let's take stock. We've pointed out that beliefs have two central functional roles—appraisal and guidance—and that this architecture is shared with other "emotion-type" mental states. From there, we detoured into a mechanistically detailed analyses of how emotion-type states instantiate their guidance function, which sheds light on what kind of control we can (and cannot) exert over this function.

The hope was that this perspective would offer insights applicable to beliefs, given the noted architectural similarity.[20] On the appraisal side, both emotions and belief states form and update automatically and nondeliberatively in response to state-appropriate stimuli. Just as threat stimuli automatically elicit fear, evidence (appraised as such) automatically elicits the formation/updating of belief states. In both cases, we are limited in our ability to intervene on the front-end: state-formation is restricted by responsiveness to state-relevant input.

Of primary interest to us here is the similarity on the guidance side: once elicited, beliefs and emotions both automatically affect a diverse range of psychological mechanisms in state-congruent ways. Indeed, both seem to influence the *same* wide range of cognitive mechanisms and processes. Like emotions, belief states affect our patterns of attention (guiding it towards information that is relevant (Shinoda et al., 2001), supportive of important beliefs (Rajsic et al., 2015), or surprisingly incongruent (Võ & Henderson, 2009)); what we encode into memory and how we recall that information (e.g., Frost et al., 2015; Tuckey & Brewer, 2003; Brewer & Treyens, 1981; Pezdek et al., 1989), how we select actions and set goals (e.g., though shaping assessments about what options are possible (Phillips et al., 2019; Phillips & Cushman, 2017) and what goals we should stick with (Kushnir, forthcoming; Cushman & Morris, 2015)), what thoughts and options spontaneously come to mind (e.g., Bear et al., 2020; Mills & Phillips, 2022), how we evaluate novel information and draw inferences—and so on, across a range of psychological processes. (These examples are intended as illustrative, not exhaustive.) If beliefs share the appraisal-guidance architecture of emotion-type states,

[20] Some recent accounts have even argued explicitly that beliefs *are* affective or emotional states (Railton, 2014; McCormick, 2022).

and through their guidance function exert influence over the same wide range of psychological processes as emotion-type states, it's reasonable to infer that the guidance processes of emotion-type states and beliefs are mechanistically similar: that beliefs also shape cognition and action via the production of state-congruent biases across diverse cognitive mechanisms.

If all that is plausible, this gets us to the matter of control. If beliefs cause default state-consistent effects across a range of psychological mechanisms in the same way that emotion-type states do, then we ought to be able to control belief's guiding function in the same way: via sequences of cognitive control actions deployed to block those default state-congruent effects across mechanisms. We can prevent belief-appraisals from instantiating their default effects on patterns of attention, thought, memory, goal selection, planning, reasoning, deliberation, action, and so on—systematically blocking the guiding function, and redirecting default responses towards motivation-congruent patterns. Moreover, we ought to be able to deploy this back-end control for any set of moral, practical, or otherwise goal-directed reasons, in cases where our appraisal states are inconsistent with our practical goals or motivations and we don't want them to have their usual effects on our patterns of reasoning, cognition, and action. This reveals an important kind of control: even if our (direct) influence on appraisal side of belief is constrained by evidential reasons, our (direct) influence on guidance-instantiation is not.

6.1. Existing Evidence

I've motivated the possibility of back-end doxastic control via functional analogy. One might wonder: is there any empirical evidence for the claim, that we can (and do) deploy cognitive control mechanisms to regulate belief-guidance?

There has not been systematic empirical investigation into belief regulation as such, in the way it's framed here. Nonetheless, there are scattered examples of psychological phenomena that are demonstrative of something like back-end doxastic control via cognitive control. I'll briefly discuss four illustrative examples from various domains of psychology.

Cognitive Control & Lying: If there is a default bias towards belief-consistent reasoning, speech, acts, and so on, it stands to reason that telling a lie ought to involve overriding a default belief-

congruent truth response. Supporting this, studies have found that neural patterns associated with cognitive control and inhibition mechanisms are activated in lying (but not honest reporting; Nuñez et al., 2005; Ofen et al., 2017; Vartanian et al., 2012; Yin et al., 2016). Relatedly, developmental research shows that children's lying ability is predicted by their executive functioning capacity (Evans & Lee, 2013).

Biased Belief Inhibition: Sometimes, successful logical reasoning requires inhibiting salient background beliefs. In De Neys and Franssens (2009), participants judged the validity of logical syllogisms. In some trials, there was a conflict between the logically correct answer and prior beliefs (e.g., "All flowers are plants. Roses are plants. Therefore, roses are flowers." is logically invalid, but has a believed conclusion); other trials had no conflict. Results showed higher error rates and slower response times for trials where there was a conflict between belief in the conclusion and the validity of the syllogism: these are classic markers of cognitive control, suggesting that in conflict trials, accurate logical reasoning requires suppressing prior belief. Moreover, when participants completed lexical decision tasks (judging whether a string of letters is a word) after the syllogisms, where the word trials either contained words that were relevant or irrelevant to the syllogism presented prior (e.g., ROSE or PEN), performance was slower specifically on relevant words that were presented after conflict trials—suggesting that participants successfully inhibited the belief in order to follow task instructions. Further studies (De Neys & Van Gelder, 2009) showed age effects corresponding with age trends in inhibitory control capacities.

Automatic vs. Controlled Prejudice: Classic social psychology of prejudice distinguishes between automatic and controlled components of these processes. A key idea is that people share background knowledge of social stereotypes, but differ in their motivation to control whether those beliefs result in prejudiced reasoning and action:[21] some systematically regulate their stereotype beliefs, while others don't. Supporting this, studies show that though low prejudice individuals normally skillfully block stereotypes from affecting their judgments and decisions, cognitive load manipulations—which

[21] There is a question about whether these stereotypes should be categorized as beliefs proper; for now, they are sufficiently belief-like.

classically impede cognitive control—interfere with their capacity to do this, thus reducing the difference between high- and low- prejudice people (Devine, 1989; Devine & Monteith, 1999; Devine & Sharp, 2009).[22] This interference effect is evidence that successfully regulating these stereotypical beliefs involves cognitive control, which people who are generally motivated to suppress stereotypes can (under normal circumstances) skillfully deploy to regulate those undesirable beliefs.

Day-Night Task. The Day-Night Task is a conflict-style task commonly used with children (Gerstadt et al., 1994; Montgomery & Koeltzow, 2010). Children are presented with cards showing either a nighttime sky (stars and moon) or a daytime sky (sun and blue sky), and instructed to label the daytime scene as "night" and the nighttime scene as "day." A reasonable gloss on this setup is that children form a perceptual belief about the presented stimulus ("that's a daytime sky")—but to follow task instructions, they must suppress the belief-congruent label. The Day-Night task is widely used in to investigate the development of inhibitory control and executive function (e.g., McAuley et al., 2011; Montgomery et al., 2008; Montgomery & Koeltzow, 2010), providing evidence that as children develop these cognitive capacities, they become better at suppressing default belief responses that are goal-incongruent.

A reasonable gloss on each of these tasks is that in some fashion, they involve a participant suppressing a default belief-response and redirecting the mechanism towards a task-appropriate response—and each offers evidence that cognitive control mechanisms are involved in doing so. Each is thus suggestive of a small, isolated version of the phenomenon we're after. Of course, these control exercises are highly contained, for a particular trial within a particular experimental context; accordingly, we're not at all inclined in these contexts to describe the agent as not believing that the stimulus is as it's presented. The only goal of presenting these examples is to lend plausibility to the claim that cognitive control mechanisms can be used to regulate belief states' guidance—this gives us the basic pieces we need to scale up.

[22] Related work also suggested that participants high in internal motivation to control prejudices are more successful at this kind of stereotype regulation than those with low motivation (Gordijn et al., 2004; Monteith et al., 1998). Motivation is a key component to successful cognitive control.

6.2 Scaling Up: Back to Voluntarism

We can imagine that when this belief regulation is scaled up—when an agent systematically deploys these cognitive control mechanisms against the same belief state across a wide range of psychological mechanisms and contexts, over an extended period of time, due to a sustained commitment—one could achieve a very thorough suppression of the underlying appraisal state. If an appraisal-state was inconsistent with an agent's standing goals, commitments, or other motivations, she could systematically prevent it from instantiating its guidance function: blocking its usual influence on reasoning, cognition, and action, and redirecting those mechanisms towards motivation-congruent alternatives. This moves us from the mechanistic details to the theoretical proposal: that back-end control can be deployed to instantiate a substantive form of doxastic voluntarism. Earlier, we noted that many theories conceive of both appraisal *and guidance* are constitutive functions of belief—that for something to count as a belief state, it needs to shape of reasoning, cognition, and action. Now, we have a story about how—via cognitive control mechanisms deployed against a belief state's default effects on various psychological mechanisms—an agent could systematically suppress the default effects of an appraisal state, thereby preventing the guidance function from being instantiated. With this, we seem to be taking significant steps down the path towards voluntarism.

But we're not there quite yet. Not every exercise of back-end doxastic control seems like an exercise of voluntarism. We often block belief-guidance temporarily, as in mundane activities like lying or hypothetical reasoning, without thereby losing the belief. We thus need some way of distinguishing exercises of back-end control that could count as genuinely voluntaristic from those which don't.[23] I'll propose two key features that set apart exercises of back-end doxastic control that amount to exercises of voluntarism from those that do not: the scope of the control exercises, and their motivational and normative profile.

First, in our hunt for voluntarism, our interest will be in cases where the guidance-suppression is exceptionally **comprehensive**: deployed consistently across time, across contexts, and across the

[23] Thanks to an anonymous referee for emphasizing this crucial point.

full suite of psychological and behavioral mechanisms. Many mundane exercises of belief suppression are temporary, deployed for short amounts of time or in isolated contexts (such as supposing for an argument), or target only a small subset of the psychological or behavioral mechanisms influenced by the appraisal (such as in telling lies to particular people). Only when back-end control is highly systematic—such that the default guidance function is thoroughly and reliably blocked in both thought and action, consistently over time and across contexts—does that it begin to threaten the appraisal's status as a genuine belief state. This is not to say that perfect back-end control must be achieved; it is part of the cognitive control apparatus that these effortful interventions are prone to failures (see §8). But the control efforts must be targeted comprehensively, and presumably hit some level of sufficient success and reliability. Notably, an agent who comprehensively deploys back-end control over contexts, time, and mechanisms is also likely to become increasingly effective in their control efforts, as this appraisal-suppression becomes increasingly skilled and habitual.

Second, the cases of interest will be characterized by a **self-directed commitment** to the systematic back-end intervention. Even a lie could be very thorough: someone could deceive everyone she encounters (though the control acts involved might still only target outward behavioral/speech patterns). But when an appraisal is seen by an agent as inconsistent with her goals or commitments in an internal, self-directed way, she won't just override it in her outward interactions with others; instead, she will intervene even in her own inner life and private mind, where she is subject only to her own personal standards. When the guidance intervention is self-directed in this way, the agent will no longer endorse the appraisal state in the way we normally think of agents as endorsing their beliefs: she no longer endorses the appraisal *as* a source of guidance, and this is what motivates the back-end intervention. Normatively, she might hold herself to account not for her appraisal of the evidence, but instead to the alternative state achieved through the guidance-intervention; *that* is the belief state with which she identifies and perhaps even ascribes to herself—that is what she is committed to reasoning, thinking, and acting on the basis of.[24]

[24] Could this mean the agent is, in some sense, lying to *herself*? Plausibly. I think the mechanisms described here have promise for understanding self-deception; exploring that is a task for future work.

We've now built up our profile of interest: the agent has an evidence-appraisal that *p*, but that *p*-appraisal stands in some kind of conflict with the agent's standing commitments or motivations. She thus refuses to endorse this appraisal, systematically blocking its characteristic guiding effects—not letting it shape the way she thinks, reasons, and acts, and redirecting default responses towards patterns that better align with her commitments. In so doing, she thoroughly cuts of the guidance function of the *p*-appraisal.

Recall a key point from our introduction of the functional roles of belief: many philosophers think guidance is a crucial part of the story for what makes something a *belief state*. Some hold two-pronged views that take both appraisal and guidance to be definitional of believing. Others hold one-pronged views that only guidance matters for characterizing belief. Either way, the crucial point is that for someone to really believe *p*, *p* needs to instantiate the guidance function: the agent's patterns of thinking and acting must actually be shaped in *p*-congruent ways. But our cases of interest, that doesn't happen: guidance is not instantiated, due to the agent's deployment of back-end doxastic control. To the extent that we take on board the functionalist idea that for something to be a mental state of a particular kind it must actually play the role(s) characteristic of that state, then exercises of systematic back-end doxastic control *thereby* enable an agent to prevent an appraisal from becoming a genuine belief. The appraisal is no longer doing the thing that belief states do—guiding reasoning, cognition, and action—and so, it's not *really* a belief state.

Finally, we can state the core proposal:

> **Back-End Doxastic Voluntarism:** Systematically blocking an appraisal-state from instantiating its characteristic guidance function—via extended, skilled sequences of cognitive control acts deployed across a full range of psychological and behavioral mechanisms, consistently across contexts and time, and motivated by a self-directed commitment not to be guided by the appraisal—is an exercise of doxastic voluntarism.

In other words: insofar as we think—as many theories of belief do—that playing the guidance role is a necessary condition of something being a belief state, then systematic exercises of back-end doxastic control can make an agent fail to meet a necessary condition of believing despite her evidence-appraisal—such that she thereby exercises a form of doxastic voluntarism, preventing herself from

having the belief state she would otherwise have. This form of voluntarism is *direct* (the capacity to intervene on guidance is not mediated by some other process), *non-evidential* (these mechanisms can be deployed in response to goal-directed, practical, and moral reasons), and *intentional* (this control is exercised volitionally, overriding default psychological processes). Exercises of back-end voluntarism will occur when an appraisal is at odds with an agent's other deeply-held commitments or motivations; this will be necessary for the regulation to be sufficiently thoroughgoing, as having the internal motivation to override default psychological responses is a key part of the cognitive control apparatus. This form of voluntarism is not quick and easy—it involves ongoing, cognitively effortful mental work (see §8.1), and may be both psychologically and practically costly. Although mundane exercises of back-end doxastic control are ubiquitous, thoroughgoing exercises of voluntarism are probably relatively more uncommon affairs. But insofar as it's psychologically possible when an agent is appropriately motivated, it can offer a route to a significant kind of doxastic agency.

Consider a soccer player, Arleen, who rationally appraises her team's odds of making the playoffs to be low—they'll have to win some unlikely games, some key players must overcome injuries, etc. As captain, Arleen is highly motivated on behalf of her team, and she feels she needs to believe in their success despite the unfavorable odds. This motivation drives her to systematically block the default effects of her pessimistic appraisal, redirecting her patterns of thought, attention, speech, planning, and behavior towards a more optimistic outlook: this outlook is manifested in how she talks to her team, her behavior and planning regarding the season, and in all her inner habits of mind, which she cultivates with discipline and determination. Someone watching her, even with access to her internal patterns of thought and reasoning, would perceive a player highly committed to the claim that her team will make the playoffs despite the unfavorable odds—not just in what she says to others, but also through how she guides her own thinking and the standards to which she holds herself. If she does this reliably and consistently across time and contexts and even in her own mind, it no longer seems theoretically apt to describe her as believing that they won't make the playoffs—she's refusing to endorse her appraisal *as* a legitimate source of guidance. She might even describe herself as *choosing to believe* that they'll make it against all odds; she is (though she may lack insight into the mechanistic

process) instantiating this choice via the skilled deployment of cognitive control exercises to override the default guidance of her unfavorable appraisal state, redirecting her thoughts and actions in ways that are compatible with her broader motivations. The proposal of the Back-End Voluntarism thesis is that in doing all this, Arleen thereby changes her belief state—because, whatever her appraisal of the evidence, she doesn't fully instantiate the belief-syndrome if she extinguishes the guidance function.

Of course, the back and front ends of belief are not entirely independent of each other, especially over time: what's downstream at t_1 may be upstream at t_2. So as Arleen redirects her patterns of attention, memory, inquiry, action, and thought over time this may come to affect what evidence she has at her disposal and how she thinks about it—such that this process may ultimately lead to changes in her appraisal as well. Extended patterns of back-end doxastic intervention may thus double as a form of indirect front-end control, highlighting how dynamic this process is in any real case. But crucially, back-end intervention does not *reduce* to indirect front-end control. First, whether back-end intervention actually changes subsequent appraisals depends on the nature of the evidence and information available to the agent; such appraisal-change is not guaranteed. Second, the central point is that systematic back-end intervention can *itself* already be a form of voluntarism, by extinguishing guidance, even without (or before) ultimately affecting the appraisal process.

Back-end voluntarism will be available to any theory of belief that takes guidance-instantiation as constitutive of believing. This is most straightforward for theories that characterize belief only or primarily in terms of guidance, such as dispositionalist views that identify beliefs in terms of dispositions to think, reason, and act in relevant circumstances (Marcus, 1990; Schwitzgebel, 2002).Similarly for certain pragmatist theories, some of which classically tied belief to patterns of action tendencies; e.g,. Peirce writes that "different beliefs are distinguished by the different modes of action to which they give rise" (1878, p. 293). But it also works for (arguably more common) two-pronged views that think genuine belief requires both appraisal and guidance: on these views, the appraisal function means the state is belief-like in some ways, but the systematic guidance-intervention means it is *not* belief-like in other key ways. The agent thus volitionally blocks the appraisal from being

a full-blown belief state, on these views—and on any view that takes guidance to be a necessary condition or constitutive feature of believing— thus agentially changing what she believes.

I can't hope to thoroughly analyze the many different conceptions of belief and what they would say about back-end voluntarism here. But in addition to the functionalist and dispositionalist perspectives that have already been highlighted, there are many other operative theories of belief that might be friendly to back-end voluntarism. Potential examples include: interpretationalist views that characterize belief ascription in terms of the success of explaining a complex system's behavioral patterns via the intentional stance (Dennett, 1989, 1989); views in the ethics of belief that talk about belief as commitments or an agent's take on the world (e.g., Basu, 2019, 2023), suggesting a crucial role for an agent's motivations and values and arguably prioritizing guidance; and Aronowitz (2023)'s planning theory, according to which believing is an activity not reducible to having specific belief-representations.

Note that the more central a theory takes guidance to be, the more voluntaristic it will be—and the easier it will be to positively characterize the belief state achieved via back-end intervention. On a guidance-focused dispositionalist theory, the agent's positive belief state will be characterized in terms of how she redirects her patterns of reasoning, cognition, and action. The story is slightly tricker for the two-pronged functionalist: in exerting back-end control, an agent can block an appraisal that would otherwise be a full-blown belief state from being that state—but the appraisal function is still in tact. The present framework thus captures that in cases where appraisal and guidance—which are normally tightly linked—dissociate, it may be tricky for the two-pronged functionalist to characterize exactly what the agent believes. But this framework also explains why this uneasiness is the *right* diagnosis for such views: precisely because such an agent instantiates one characteristic component of believing, but not the other.

Finally, I'll note that there remains an important question about the scope of this voluntarism proposal: what kinds of belief states can we successfully deploy comprehensive back-end control against? I will leave addressing this as a project for the future, though it seems unlikely that an agent can fully succeed in deploying back-end control against *any* kind of belief state—for instance,

thoroughly regulating basic perceptional beliefs may be unlikely to succeed. There may also be an accompanying question about what kind of *motivational* profile is more or less likely to support successful back-end voluntarism. The proposal, for now, is more modest: thoroughgoing back-end control is something we *can* deploy for regulating belief states; it is a future (perhaps empirical) project to identify what features of particular belief states make back-end control more or less likely to be successful.

6.3. Appraisal without Guidance

My argument is that thoroughly blocking an appraisal state's guidance function can itself be an exercise of voluntarism—even if the appraisal function is left untouched. I thus have not challenged standard observations about the limits of our control over belief appraisal; indeed, the comparison to emotion-type states further highlights that these classic views track something importantly right about the constraints on the appraisal function. Instead, we've challenged an implicit background assumption in the voluntarism literature[25]—one which is not supported by broader philosophical perspectives on belief—that only appraisal matters for the identity of belief states, and so, that intervening on the appraisal side is the only route to voluntarism. But beliefs aren't just appraisals; they are also "the maps by which we steer" (Ramsey, 1931)—and in exerting thoroughgoing back-end control, we can, so to speak, take back the wheel.

This argument will not convince someone who insists on an appraisal-only theory of belief. For those who characterize guidance-instantiation as *only* the downstream effects of belief rather than *part of* believing, the mechanisms here are not a route to voluntarism—they must be classified as something else (perhaps something like acceptance).[26] But although a full argument against appraisal-only conceptions of belief is beyond the scope of this paper, I'll briefly note that accepting such

[25] An exception is Ginet (2001); see §7.1.

[26] Specifically, some notions of acceptance, such as Bratman's (1992), characterize acceptance in terms of an agent *departing* from their default cognitive background of belief.

theories is not without cost.[27] There is the already-discussed fact that there is a strong tradition in philosophy of mind of emphasizing the importance of guidance; appraisal-only views lose the ability to capture this part of the commonly endorsed psychofunctionalist viewpoint. Beyond this, conceiving of guidance as constitutive of also belief offers continuity with how cognitive scientists think of other mental states as well; an emotion, for instance, is thought of as not merely the appraisal, but the appraisal *coupled with* the its expression across various cognitive and behavioral processes (Sripada, 2021; Scherer, 2022). An emotion-appraisal (e.g., an assessment of threat) that led to *no* discernable guidance effects (e.g., no shaping of patterns of thought or behavior) would not be thought of as a full instance of an emotion (e.g., fear).

Another way to appreciate the importance of guidance is to consider a case where guidance is not instantiated due to non-volitional forces. Imagine Connor has an evidence-appraisal that Minnesotans are friendly—but also has a highly odd brain tumor that selectively prevents this appraisal from ever actually shaping her thinking, reasoning, and action. (We'll set aside our aspirations of empirical plausibility for the moment.) Although she takes her evidence to support the friendliness of Minnesotans, in all relevant contexts she finds herself inclined towards reasoning and acting against this appraisal—the tumor causes her to react with caution around the Minnesotans she encounters, assume the worst of them and treat them with distrust, and to be constantly wary when in Minneapolis. Does Connor really believe that Minnesotans are friendly? It likely strikes many of us as quite uncomfortable to say that she does—precisely because the tumor prevents the instantiation of the full belief-syndrome. In this case, the intervention on guidance comes from an (agentially) exogenous

[27] There's an interesting question about how to classify representationalist theories of belief within this framework. As noted in §3, representationalists often appeal to both front-end and back-end functions, and some accounts (e.g., Mandelbaum and Porot, 2021; Quilty-Dunn and Mandelbaum, 2018) are explicitly psychofunctionalist accounts. However, it's possible that when pressed, though, some representationalists would say that the belief is really the stored representation, and that the importance of the back-end is just that a representation of the right type to enter into appropriate computational processes—regardless of whether it actually does. Back-end control might thus not be genuine voluntarism for this kind of representationalist.

force. My proposal, of course, is that we can exert this kind of intervention intentionally, via back-end doxastic control—and thereby exercise a form of voluntarism.

The present argument can thus be read as conditional—*insofar as* guidance-instantiation is constitutive of believing, back-end control offers a route to voluntarism—coupled with the observation that the antecedent is widely accepted, and with good reason. But a serious attempt to settle disagreement about the antecedent must be saved for another time.

7. Two More "Is It Really Voluntarism?" Objections

I'll consider two further objections to the Back-End Voluntarism thesis: first a worry about which dispositions count for the dispositionalist, and the second about whether the account gets us not belief but some adjacent doxastic attitude, like suspension of judgment.

7.1. Which Dispositions?

I've argued that back-end control can amount to voluntarism on dispositionalist theories of belief, among others. But I've also noted that on the cognitive control framework, an agent cannot directly change the fact *that* a particular response is the default response, only whether default response is actualized. Perhaps these points are in tension: the dispositionalist might press that I haven't defended genuine voluntarism even on their account, because you're not changing that you *have a disposition* to think, reason, act, and so on in particular ways by default—you're just preventing its actualization.

Whether comprehensive back-end control counts as controlling the belief-disposition ultimately depends on precisely how we flesh out the nature of these belief-dispositions. There are (at least) two ways to do this: in terms of the low-level default responses pulses; or in terms of whether an agent systematically and reliably regulates those low-level default responses pulses. In either case the agent is disposed in a particular way—there's a sense in which she is cognitively disposed towards certain default patterns, but there's also a sense in which she's disposed towards a different set of psychological and behavioral patterns given her motivation to thoroughly and skillfully regulate those default response pulses. I think it's just not obvious which kind of disposition counts as the belief-disposition.

Consider, for analogy, a skilled archer. When preparing to take a shot, she must stay highly focused on her body position, breath, aim, and target. Suppose that while she is preparing her shot, a chaotic flock of birds bursts through the trees nearby. Humans have a default cognitive tendency to shift their gaze towards objects moving incongruently in the visual field. But our experienced archer reliably and skillfully overrides that default tendency and stays focused on her target (using cognitive control mechanisms). It seems that there is both a sense in which she is disposed to look up towards the birds, and also a sense in which she is disposed to stay focused on the target. Further, it seems non-obvious whether we should say that her "real" attentional disposition is the basic cognitive one, or the skilled, effortful focused one—that depends on our explanatory target.

The same, I suggest, holds for belief, when an agent skillfully and systematically overrides the appraisal-driven responses pulses. Because existing accounts have not typically analyzed belief as a prepotent mental state that generates responses pulses across diverse mechanisms, it's an open question how dispositionalists would characterize these competing considerations. If we look at existing dispositionalist accounts with an eye to this question, it is plausible—though not in every case settled—that an agent who deploys comprehensive back-end doxastic control counts as believing.

I'll briefly highlight three examples. First, Schwitzgebel (2002) characterizes belief as a cluster of behavioral, phenomenal, and cognitive dispositions stereotypical of the belief, which the agent would manifest *ceterus paribus*, across a wide and important range of circumstances. An agent committed to back-end doxastic regulation might, across a wide and important range of circumstances, reliably intervene on a belief state's guiding role across these domains. Second, Zimmerman (2018) develops his pragmatist account of belief in terms of dispositions to reason, think, and act *when our attention and self-control* are brought to bear on matters, rather than those dispositions that guide us automatically. He writes: "to believe something at a given time is to be do disposed that you would use that information to guide those relatively attentive and self-controlled activities you might engage in" (p. 1), and goes on to explicitly contrast "degree of assimilation"—roughly, the degree to which we are automatically guided by some piece of information—from belief, because belief sometimes involves overriding automatically assimilated reactions (pp. 2-3). An agent deploying thorough back-end

control would likely satisfy his account of believing. Third, Ginet (2001) characterizes belief as a disposition to act and rely on *p* in various situations; this involves things like staking something on *p*, counting on *p*, or not planning for the possibility that not-*p* (pp. 66-67). Ginet's account is particularly relevant because his (2001)'s primary goal is to offer a defense of doxastic voluntarism. He argues that, given this dispositional conception of belief, we can (in at least some cases) decide whether to believe *p* by (for example) not preparing oneself for the possibility that not-*p*. Ginet develops this idea in strikingly cognitive terms, noting that "in the right circumstances, it *can take effort to avoid preparing oneself* for the possibility that not-*p*… to *suppress considering that possibility and what to do if it is realized*" (2001, p. 66; emphasis added). He even describes this in terms of "resisting an impulse"—such as choosing not to believe someone was injured in a car crash by *suppressing all impulse* to imagine her possible injuries or plan how he will handle it if she were injured.

Each of these accounts plausibly could (or for Ginet, nearly explicitly does) diagnose an agent who skillfully and systematically deploys back-end doxastic control as having a disposition that qualifies as *believing*—and thus, that back-end control counts as controlling what one believes. I won't take a definitive stand here on how dispositionalist accounts *should* incorporate back-end control into their frameworks; I am simply pointing out that that the kind of thing philosophers have in mind when they define beliefs in terms of dispositions to reason, think, and may well be something we can control via thorough comprehensive, committed back-end doxastic regulation—and thus, that it remains plausible to think of this as genuine voluntarism on these accounts.

7.2. Belief, or Something Nearby?

Some recent work has argued that even if we cannot believe at will, we can *suspend judgment* at will. Perhaps what I have really done is offer a mechanistic account of suspension?[28]

Suspension of judgment can mean different things, though accounts are generally unified in positing some kind of committed neutrality towards the target proposition. This neutrality could be manifesting an "inquiring attitude" (Friedman, 2013, 2017); refraining from judgment about a matter

[28] Thanks to an anonymous referee for encouraging me to consider this.

(McGrath, 2021; Ross, 2022); or taking up an intermediate degree of confidence (Fritz, 2021).[29] The mechanisms of back-end doxastic control elucidated here very plausibly ultimately underlie some conceptions of suspension: specifically, notions that hold suspension can be willfully adopted even when an agent's confidence in the target proposition is quite high (or low), but the agent nonetheless wants to maintain a neutral attitude.

However, the present account is not *only* an account of suspension. First, though back-end doxastic control can surely be deployed because of an agent's commitment to an attitude of neutrality, this is not the only possible motivational story. An agent might be committed to the outright denial of her appraisal, or to boosting her commitment to the proposition beyond the confidence licensed by the appraisal. Recall our soccer captain Arleen. She is not neutral about whether her team will make the playoffs, and not suspending judgment about their chances; her self-directed commitment is that they *do* have a real chance. If the agent's motivation for back-end intervention is outright denial or positive commitment to an alternative, describing her as suspending judgment will not be apt. Second (and relatedly), recall the earlier note that these back-end control mechanisms can be deployed against any kind of underlying appraisal state—including uncertainty. An agent who intervenes on the guidance of an uncertainty-appraisal will presumably not be committed to the neutrality characteristic of suspension: instead, she will block patterns of thought and behavior that flow from the uncertainty-appraisal and boost those mechanisms towards the claim she is committing herself to, overriding that default uncertainty. The mechanisms of interest, in other words, don't just allow for a negative suppression: as the default appraisal-responses are redirected, they allow for manifestation of commitment to some positive alternative.

This suggests that how we'll ultimately want to describe an agent who deploys systematic back-end doxastic control depends on a combination of her broader motivational profile, the nature of the underlying appraisal state, and how she redirects her default patterns of cognition and behavior. Some

[29] McGrath (2021) offers alternative labels for each of these, and maps their relations.

cases may aptly be described as suspension of judgment, but others will not.[30] Importantly, the key point holds regardless: that, although an agent cannot will herself directly into an alternative appraisal against her evidence, she can deploy back-end doxastic control to instantiate a doxastic state that departs quite significantly from her appraisal.

8. Further Upshots

I'll close by highlighting some upshots of this new voluntaristic framework.

8.1. A Distinctive Psychological Profile

One significant upshot of adopting the cognitive control framework is that it reveals various psychological features which we might not have been accustomed to thinking about in the context of doxastic control.

For example: mental regulation via cognitive control is characteristically effortful, involving executive functioning skills to identify and override default responses, and this effort is phenomenologically appreciable. Back-end doxastic control will thus often feel cognitively effortful: instantiating this form of voluntarism takes cognitive work. Another example concerns individual differences in capacity, skill, and habit. People vary substantially in their general capacity for cognitive control regulation (within "normal" and disordered ranges).[31] Moreover, it is familiar people's ability to successfully regulate other emotion-type can become more skilled or habitual with practice; we might expect the same for doxastic states. Regulating an unwanted appraisal may be more difficult at first, or for someone not used to or disciplined in this kind of cognitive regulation, or for someone low in general cognitive control capacity. Another relevant factor in determining success will be the strength of the agent's motivation,[32] which can also vary across people and contexts. All this raises intriguing questions regarding the possibility of substantial individual and contextual differences in

[30] (Some notions of) suspension may also be necessarily temporary, undertaken for the purpose of inquiry (Friedman's view) or because an agent expects better evidence in the future (see McGrath).

[31] Cognitive control is frequently measured as an individual difference measure; this assumption is ubiquitous (e.g., Duckworth & Kern, 2011; Weigard et al., 2021; Weigard & Sripada, 2021). These references highlight some recent work using computational approaches to study these individual differences (Weigard's work), and the relationship between various self-control measures (Duckworth's work).

[32] Frömer et al. (2021) discuss the importance of motivation and expected reward in cognitive control allocation.

people's capacity to instantiate back-end doxastic voluntarism—an idea that has not been thoroughly considered in the context of voluntarism discussions.

Another upshot concerns the broader framework: back-end doxastic control turns out be an exercise of *self-control*, just as Sripada (2021) characterizes control of emotion-type states. Back-end control is instantiated via an extended series of basic mental acts deployed across mechanisms and over time: it involves the ongoing regulation of default psychological tendencies in response to sustained practical motivations, goals, or commitments. This is a different kind of action profile than we might have expected when thinking about voluntarism, and opens up new room for theorizing about doxastic self-control—including other kinds of doxastic self-control strategies. For instance, Bermúdez (2021) argues that skilled self-control involves selecting diachronic strategies (such as avoiding particular environments or stimuli) that reduce the need to deploy cognitively difficult synchronic suppression strategies. Translating this framework into the context of belief may enable us to more systematically unite various kinds of "belief management" strategies that philosophers often point to; fleshing out this broader picture is a project for future work.

These features offer a preview of the richness of the psychological profile diagnosed by back-end voluntarism, and point to intriguing new questions (both theoretical and potentially empirical) regarding the cognitive landscape of doxastic control.

8.2. The Ethics of Belief

Back-end voluntarism also has potential upshots for debates in the ethics and pragmatics of belief. These debates often center around cases in which an agent seems to have moral or practical reason to believe against her evidence. Thus, whether one can be doxastically responsive to non-evidential reasons is central—and is often noted as a core theoretical challenge for the field, when constrained by the traditional focus on front-end involuntarism. Back-end control may give us precisely what we need: a story about how we can be thoroughly doxastically responsive to non-evidential reasons, in a way that is: expressive of an agent's values and motivations; explained in terms of independently plausible cognitive mechanisms; and compatible with orthodox defenses of involuntarism. The account thus has potential to open up new space for navigating issues in the ethics

of belief. This is not to say that ethicists of belief ought only to care about guidance; questions about the normative status of both components remain live. But when theorists move from thinking about beliefs as assessments of evidence towards an agent's take on our outlook towards the world, such conceptions begin to implicitly import more guidance-focused notions of belief than is sometimes appreciated. Indeed—though I will not defend this here—I ultimately think that some conceptual clarity can be brought to various debates by introducing the distinction between the appraisal and guidance components of belief.

Unveiling the psychological processes at play also offers opportunity for normative nuance, insofar as we want to our moral theorizing to be sensitive to the aforementioned cognitive features. How, for instance, might our ethics of belief change when we think about doxastic regulation as self-control, or when we account for substantive individual differences in doxastic control capacity or skill? These questions, and others, remain to be worked out.

9. Conclusion

I have defended three claims: (1) when thinking about doxastic control, we have been too narrowly focused on appraisal, at the cost of considering guidance; (2) we can exert systematic back-end control over belief's guidance function, via the deployment of skilled sequences of cognitive control mechanisms; and (3) when deployed comprehensively and out of a self-directed commitment, systematic back-end control can amount to a substantive kind of doxastic voluntarism. I hope to have delivered a psychologically plausible proposal about doxastic control that can shed new light on philosophical questions of doxastic agency.

By rooting this discussion in the broader cognitive science of mental control, I have also aimed at two further goals. First, I want to emphasize that these mechanisms are ones that people plausibly *actually use* to regulate an undesirable and goal-inconsistent appraisal states. This may help explain various cases in which people seem to believe things that appear to be at odds with the information they have—a puzzling pattern that appears in a wide range of cases that capture philosophers' interests, from self-deception or denial to ideologies and radical belief systems. Though these cases surely involve a complex variety of mechanisms, it's plausible that systematic back-end control is one

kind of (potentially underappreciated) component, when agents are highly motivated (e.g., for reasons of psychological comfort or group loyalty) to believe against their evidence. These possibilities, juxtaposed with the earlier discussion about the role of back-end control in the ethics of belief, highlight that these mechanisms can be used in ways that are good or bad—as is true of many cognitive capacities.

Finally, this methodological approach aims to bring the topic of doxastic control out of the armchair and into contact with empirical cognitive science. Despite centrally concerning the question of what kinds of control we do and don't have over mental states, the voluntarism debate has had remarkably little influence from psychological research; much of it centers around conceptual analyses of belief. This is, undoubtedly, an important facet of the project, in part because the nature of belief is a thorny philosophical question. Nonetheless, I hope to have shown that we might make progress on questions of doxastic control by leveraging existing cognitive scientific research on mental state control. What else might we learn about doxastic control using this methodology—what other similarities, and what differences, might there be between control of beliefs and other mental states? How might the mechanistic approach here inspire not only theoretical work, but possibly also experimental investigation? I think we have a lot to learn in this domain from the psychologists, and it's worthwhile to explore how far we can get with belief.[33]

[33] I am grateful to Chandra Sripada, Maegan Fairchild, Walter Sinnott-Armstrong, Z Quanbeck, Nico Orlandi, Aliosha Barranco Lopez, Anna Vaughn, Melanie Rosen, Casey Landers, Malte Hendrickx, Zach Barnett, Gabrielle Kerbel, and a very helpful anonymous reviewer for their comments on various drafts of this paper, as well as to Renée Jorgensen, Kirun Sankaran, Shanna Slank, Kevin O'Neill, Tamar Kushnir, Ben Eva, Reuben Stern, and Adam Waggoner for helpful discussion. Versions of this paper were presented at the 2023 Southern Society for Philosophy and Psychology, the 2023 Society for Philosophy and Psychology, the 2022 Narrow Ridge Philosophy Workshop, the National University of Singapore, the 2023 Duke/UNC Epistemology Workshop, the 2023 Marc Sanders Mentoring Workshop, and the 2024 Eastern American Philosophical Association Meeting; thanks to those audiences for their engagement, and to Mike Roche and Jake Green for conference comments. Thanks also to Jason Decker, for first getting me puzzled about doxastic voluntarism many years ago.

References

Adolphs, R., & Andler, D. (2018). Investigating Emotions as Functional States Distinct From Feelings. *Emotion Review*, *10*(3), 191–201. https://doi.org/10.1177/1754073918765662

Alston, W. P. (1988). The Deontological Conception of Epistemic Justification. *Philosophical Perspectives*, *2*, 257–299. JSTOR. https://doi.org/10.2307/2214077

Angie, A. D., Connelly, S., Waples, E. P., & Kligyte, V. (2011). The influence of discrete emotions on judgement and decision-making: A meta-analytic review. *Cognition and Emotion*, *25*(8), 1393–1422. https://doi.org/10.1080/02699931.2010.550751

Arpaly, N. (n.d.). Practical reasons to believe, epistemic reasons to act, and the baffled action theorist. *Philosophical Issues*, *n/a*(n/a). https://doi.org/10.1111/phis.12239

Basu, R. (2019). The wrongs of racist beliefs. *Philosophical Studies*, *176*(9), 2497–2515. https://doi.org/10.1007/s11098-018-1137-0

Basu, R. (2023). Morality of Belief II: Three Challenges and An Extension. *Philosophy Compass*, *18*(7), e12935. https://doi.org/10.1111/phc3.12935

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, *194*, 104057. https://doi.org/10.1016/j.cognition.2019.104057

Begby, E. (2021). *Prejudice: A Study in Non-Ideal Epistemology*. Oxford University Press.

Bennett, J. (1990). Why Is Belief Involuntary? *Analysis*, *50*(2), 87–107. https://doi.org/10.2307/3328852

Bermúdez, J. P. (2021). The skill of self-control. *Synthese*, *199*(3), 6251–6273. https://doi.org/10.1007/s11229-021-03068-w

Block, N. (1980). What is Functionalism? In N. Block (Ed.), *Readings in the Philosophy of Psychology*.

Bodenhausen, G. V., Sheppard, L. A., & Kramer, G. P. (1994). Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of Social Psychology*, *24*(1), 45–62. https://doi.org/10.1002/ejsp.2420240104

Bower, G. H. (1991). Mood Congruity of Social Judgments. In *Emotion and Social Judgments* (pp. 31–53). Garland Science. https://doi.org/10.4324/9781003058731-3

Bower, G. H., & Cohen, P. R. (1982). Emotional Influences in Memory and Thinking: Data and Theory. In *Affect and Cognition* (pp. 301–342). Psychology Press. https://doi.org/10.4324/9781315802756-21

Bratman, M. E. (1992). Practical Reasoning and Acceptance in a Context. *Mind*, *101*(401), 1–15. JSTOR.

Chrisman, M. (2008). Ought to Believe. *The Journal of Philosophy*, *105*(7), 346–370. https://doi.org/10.5840/jphil2008105736

Chrisman, M. (2022). Doxastic Involuntarism and 'Ought to Believe.' In M. Chrisman (Ed.), *Belief, Agency, and Knowledge: Essays on Epistemic Normativity* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780192898852.003.0006

Cohen, L. J. (1989). Belief and Acceptance. *Mind*, *98*(391), 367–389. JSTOR.

Cohen, L. J. (1995). *Essay on Belief and Acceptance*. Oxford University Press UK.

D'Arms, J., & Jacobson, D. (2000). The Moralistic Fallacy: On the 'Appropriateness' of Emotions. *Philosophy and Phenomenological Research*, *61*(1), 65–90.

D'Arms, J., & Jacobson, D. (2023). *Rational Sentimentalism*. Oxford University Press.

Davidson, D. (1985). Incoherence and Irrationality. *Dialectica*, *39*(4), 345–354. https://doi.org/10.1111/j.1746-8361.1985.tb01603.x

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, *113*(1), 45–61. https://doi.org/10.1016/j.cognition.2009.07.009

De Neys, W., & Van Gelder, E. (2009). Logic and belief across the lifespan: The rise and fall of belief inhibition during syllogistic reasoning. *Developmental Science*, *12*(1), 123–130. https://doi.org/10.1111/j.1467-7687.2008.00746.x

Dennett, D. C. (1989). *The Intentional Stance*. MIT Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, *56*, 5–18. https://doi.org/10.1037/0022-3514.56.1.5

Devine, P. G., & Monteith, M. J. (1999). Automaticity and control in stereotyping. In *Dual-process theories in social psychology* (pp. 339–360). The Guilford Press.

Devine, P. G., & Sharp, L. B. (2009). Automaticity and control in stereotyping and prejudice. In *Handbook of prejudice, stereotyping, and discrimination* (pp. 61–87). Psychology Press.

Dretske, F. (1991). *Explaining Behavior: Reasons in a World of Causes*. The MIT Press. https://doi.org/10.7551/mitpress/2927.001.0001

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Egan, A. (2008). Imagination, delusion, and self-deception. In *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Psychology Press.

Ellsworth, P. C. (2013). Appraisal Theory: Old and New Questions. *Emotion Review*, *5*(2), 125–131. https://doi.org/10.1177/1754073912463617

Ellsworth, P. C., & Scherer, K. R. (2003). Appraisal processes in emotion. In *Handbook of affective sciences* (pp. 572–595). Oxford University Press.

Evans, A. D., & Lee, K. (2013). Emergence of lying in very young children. *Developmental Psychology*, *49*, 1958–1963. https://doi.org/10.1037/a0031409

Eysenck, M. W., Mogg, K., May, J., Richards, A., & Mathews, A. (1991). Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of Abnormal Psychology*, *100*, 144–150. https://doi.org/10.1037/0021-843X.100.2.144

Feldman, R. (2000). The Ethics of Belief. *Philosophy and Phenomenological Research*, *60*(3), 667–695.

Flores, C. (Forthcoming). Resistant Beliefs, Responsive Believers. *The Journal of Philosophy*.

Frankish, K. (2007). Deciding to Believe Again. *Mind*, *116*(463), 523–548. https://doi.org/10.1093/mind/fzm523

Friedman, J. (2013). Suspended judgment. *Philosophical Studies*, *162*(2), 165–181. https://doi.org/10.1007/s11098-011-9753-y

Friedman, J. (2017). Why Suspend Judging? *Noûs*, *51*(2), 302–326. https://doi.org/10.1111/nous.12137

Friedman, J. (2019). Inquiry and Belief. *Noûs*, *53*(2), 296–315. https://doi.org/10.1111/nous.12222

Frijda, N. H. (1986). *The Emotions*. Cambridge University Press.

Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cognition and Emotion*, *1*(2), 115–143. https://doi.org/10.1080/02699938708408043

Frijda, N. H., Kuipers, P., & ter Schure, E. (1989). Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology*, *57*, 212–228. https://doi.org/10.1037/0022-3514.57.2.212

Fritz, J. (2021). Hope, Worry, and Suspension of Judgment. *Canadian Journal of Philosophy*, *51*(8), 573–587. https://doi.org/10.1017/can.2022.20

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21315-z

Frost, P., Casey, B., Griffin, K., Raymundo, L., Farrell, C., & Carrigan, R. (2015). The Influence of Confirmation Bias on Memory and Source Monitoring. *The Journal of General Psychology*, *142*(4), 238–252. https://doi.org/10.1080/00221309.2015.1084987

Gendler, T. S. (2008a). Alief and Belief. *The Journal of Philosophy*, *105*(10), 634–663.

Gendler, T. S. (2008b). Alief in Action (and Reaction). *Mind & Language*, *23*(5), 552–585. https://doi.org/10.1111/j.1468-0017.2008.00352.x

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 312–7 years old on a stroop- like day-night test. *Cognition*, *53*(2), 129–153. https://doi.org/10.1016/0010-0277(94)90068-X

Gordijn, E. H., Hindriks, I., Koomen, W., Dijksterhuis, A., & Van Knippenberg, A. (2004). Consequences of Stereotype Suppression and Internal Suppression Motivation: A Self-Regulation Approach. *Personality and Social Psychology Bulletin*, *30*(2), 212–224. https://doi.org/10.1177/0146167203259935

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, *5*(9), 394–400. https://doi.org/10.1016/S1364-6613(00)01707-1

Helton, G. (2020). If You Can't Change What You Believe, You Don't Believe It. *Noûs*, *54*(3), 501–526. https://doi.org/10.1111/nous.12265

Hieronymi, P. (2006). Controlling Attitudes. *Pacific Philosophical Quarterly*, *87*(1), 45–74. https://doi.org/10.1111/j.1468-0114.2006.00247.x

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, *161*(3), 357–373. https://doi.org/10.1007/s11229-006-9089-x

Jackson, E. G. (2021). A Permissivist Defense of Pascal's Wager. *Erkenntnis*. https://doi.org/10.1007/s10670-021-00454-1

Kelly, T. (2002). The Rationality of Belief and Some Other Propositional Attitudes. *Philosophical Studies*, *110*(2), 163–196. https://doi.org/10.1023/A:1020212716425

Keltner, D., & Gross, J. J. (1999). Functional Accounts of Emotions. *Cognition and Emotion*, *13*(5), 467–480. https://doi.org/10.1080/026999399379140

Kensinger, E. A. (2009). Remembering the Details: Effects of Emotion. *Emotion Review*, *1*(2), 99–113. https://doi.org/10.1177/1754073908100432

Kieval, P. H. (2022). Permission to believe is not permission to believe at will. *Synthese*, *200*(5), 347. https://doi.org/10.1007/s11229-022-03845-1

Kushnir, T. (n.d.). *Developmental Pathways to Possibility Beliefs*.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, *14*(4), 473–493. https://doi.org/10.1080/026999300402763

Levin, J. (2021). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2021/entries/functionalism/

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements: Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs*, *49*(4), 800–823. https://doi.org/10.1111/nous.12074

Levy, N., & Mandelbaum, E. (2014). The Powers that Bind: Doxastic Voluntarism and Epistemic Obligation. In J. Matheson & R. Vitz (Eds.), *The Ethics of Belief: Individual and Social* (pp. 15–32). Oxford University Press.

Lewis, D. (1974). Radical Interpretation. *Synthese*, *27*(3/4), 331–344.

Madva, A. (2016). Why implicit attitudes are (probably) not beliefs. *Synthese*, *193*(8), 2659–2684. https://doi.org/10.1007/s11229-015-0874-2

Mandelbaum, E. (2016). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias: Attitude, Inference, Association. *Noûs*, *50*(3), 629–658. https://doi.org/10.1111/nous.12089

Marcus, R. B. (1990). Some Revisionary Proposals about Belief and Believing. *Philosophy and Phenomenological Research*, *50*, 133–153. https://doi.org/10.2307/2108036

McAuley, T., Christ, S. E., & White, D. A. (2011). Mapping the Development of Response Inhibition in Young Children Using a Modified Day-Night Task. *Developmental Neuropsychology*, *36*(5), 539–551. https://doi.org/10.1080/87565641.2010.549871

McGrath, M. (2021). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*, *55*(2), 463–484. https://doi.org/10.1111/nous.12323

Mills, T., & Phillips, J. S. (2022). *Locating what comes to mind in empirically derived representational spaces*. PsyArXiv. https://doi.org/10.31234/osf.io/fds9q

Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a Stereotype Control Strategy. *Personality and Social Psychology Review*, *2*(1), 63–82. https://doi.org/10.1207/s15327957pspr0201_4

Montgomery, D. E., Anderson, M., & Uhl, E. (2008). Interference control in preschoolers: Factors influencing performance on the day–night task. *Infant and Child Development*, *17*(5), 457–470. https://doi.org/10.1002/icd.559

Montgomery, D. E., & Koeltzow, T. E. (2010). A review of the day–night task: The Stroop paradigm and interference control in young children. *Developmental Review*, *30*(3), 308–330. https://doi.org/10.1016/j.dr.2010.07.001

Moors, A. (2014). Flavors of Appraisal Theories of Emotion. *Emotion Review*, *6*(4), 303–307. https://doi.org/10.1177/1754073914534477

Nolfi, K. (2015). How to be a Normativist about the Nature of Belief. *Pacific Philosophical Quarterly*, *96*(2), 181–204. https://doi.org/10.1111/papq.12071

Nuñez, J. M., Casey, B. J., Egner, T., Hare, T., & Hirsch, J. (2005). Intentional false responding shares neural substrates with response conflict and cognitive control. *NeuroImage*, *25*(1), 267–277. https://doi.org/10.1016/j.neuroimage.2004.10.041

Ofen, N., Whitfield-Gabrieli, S., Chai, X. J., Schwarzlose, R. F., & Gabrieli, J. D. E. (2017). Neural correlates of deception: Lying about past events and personal beliefs. *Social Cognitive and Affective Neuroscience*, *12*(1), 116–127. https://doi.org/10.1093/scan/nsw151

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General*, *130*, 466–478. https://doi.org/10.1037/0096-3445.130.3.466

Peirce, C. (1878). How to Make our Ideas Clear. *Popular Science Monthly*, 286–302.

Phelps, E. A., Ling, S., & Carrasco, M. (2006). Emotion Facilitates Perception and Potentiates the Perceptual Benefits of Attention. *Psychological Science*, *17*(4), 292–299. https://doi.org/10.1111/j.1467-9280.2006.01701.x

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1619717114

Phillips, J., Morris, A., & Cushman, F. (2019). How We Know What Not To Think. *Trends in Cognitive Sciences*, *23*(12), 1026–1040. https://doi.org/10.1016/j.tics.2019.09.007

Porot, N., & Mandelbaum, E. (2021). The science of belief: A progress report. *WIREs Cognitive Science*, *12*(2), e1539. https://doi.org/10.1002/wcs.1539

Quanbeck, Z., & Worsnip, A. (n.d.). A Permissivist Alternative to Encroachment. *Philosophers' Imprint*.

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, *175*(9), 2353–2372. https://doi.org/10.1007/s11098-017-0962-x

Railton, P. (2014). Reliance, Trust, and Belief. *Inquiry*, *57*(1), 122–150. https://doi.org/10.1080/0020174X.2014.858419

Ramsey, F. (1931). *The Foundations of Mathematics and Other Logical Essays*. Routledge Kegan & Paul.

Roeber, B. (2019). Evidence, Judgment, and Belief at Will. *Mind*, *128*(511), 837–859. https://doi.org/10.1093/mind/fzy065

Roeber, B. (2020). Permissive Situations and Direct Doxastic Control. *Philosophy and Phenomenological Research*, *101*(2), 415–431. https://doi.org/10.1111/phpr.12594

Roseman, I. J., & Smith, C. A. (2001). Appraisal Theory: Overview, Assumptions, Varieties, Controversies. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research* (pp. 3–19). Oxford University Press.

Ross, L. (2022). Profiling, Neutrality, and Social Equality. *Australasian Journal of Philosophy*, *100*(4), 808–824. https://doi.org/10.1080/00048402.2021.1926522

Ryan, S. (2003). Doxastic Compatibilism and the Ethics of Belief. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *114*(1/2), 47–79. JSTOR.

Ryle, G. (1949). *The Concept of Mind*. Barnes and Noble. https://doi.org/10.4324/9780203875858

Scarantino, A. (2014). The motivational theory of emotions. In *Moral psychology and human agency: Philosophical essays on the science of ethics* (pp. 156–185). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198717812.003.0008

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. *Noûs*, *36*(2), 249–275. https://doi.org/10.1111/1468-0068.00370

Schwitzgebel, E. (2006). Belief. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/belief/?TB_iframe=true&width=370.8&height=658.8#WhatItBeli

Schwitzgebel, E. (2011). Belief. In *The Routledge Companion to Epistemology*. Routledge.

Scott-Kakures, D. (1994). On Belief and the Captivity of the Will. *Philosophy and Phenomenological Research*, *54*(1), 77–103.

Shah, N. (2002, August 1). *Clearing Space For Doxastic Voluntarism*. The Monist. https://doi.org/10.5840/monist200285326

Shah, N. (2003). How Truth Governs Belief. *The Philosophical Review*, *112*(4), 447–482.

Shah, N. (2006). A New Argument for Evidentialism. *The Philosophical Quarterly*, *56*(225), 481–498. https://doi.org/10.1111/j.1467-9213.2006.454.x

Shah, N., & Velleman, J. D. (2005a). Doxastic Deliberation. *The Philosophical Review*, *114*(4), 497–534.

Shah, N., & Velleman, J. D. (2005b). Doxastic Deliberation. *The Philosophical Review*, *114*(4), 497–534.

Singh, K. (Forthcoming). Belief as Commitment to the Truth. In E. Schwitzgebel & J. Jong (Eds.), *The Nature of Belief*. Oxford University PRess.

Smallwood, J., Fitzgerald, A., Miles, L. K., & Phillips, L. H. (2009). Shifting moods, wandering minds: Negative moods lead the mind to wander. *Emotion*, *9*, 271–276. https://doi.org/10.1037/a0014855

Sripada, C. (2021). The atoms of self-control. *Noûs*, *55*(4), 800–824. https://doi.org/10.1111/nous.12332

Steup, M. (2012). Belief control and intentionality. *Synthese*, *188*(2), 145–163. https://doi.org/10.1007/s11229-011-9919-3

Steup, M. (2017). Believing intentionally. *Synthese*, *194*(8), 2673–2694. https://doi.org/10.1007/s11229-015-0780-7

Steup, M. (2018). Doxastic Voluntarism and Up-To-Me-Ness. *International Journal of Philosophical Studies*, *26*(4), 611–618. https://doi.org/10.1080/09672559.2018.1511148

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/h0054651

Sylvan, K. (2016). The illusion of discretion. *Synthese*, *193*(6), 1635–1665. https://doi.org/10.1007/s11229-015-0796-z

Vartanian, O., Kwantes, P., & Mandel, D. R. (2012). Lying in the scanner: Localized inhibition predicts lying skill. *Neuroscience Letters*, *529*(1), 18–22. https://doi.org/10.1016/j.neulet.2012.09.019

Velleman, D. (2000). On the aim of belief. In *The Possibility of Practical Reason* (pp. 244–281). Oxford University Press.

Velleman, J. D. (2014). The Possibility of Practical Reason, 2nd Edition. *Maize Books*. https://doi.org/10.3998/maize.13240734.0001.001

Vermaire, M. (2022). In search of doxastic involuntarism. *Philosophical Studies*, *179*(2), 615–631. https://doi.org/10.1007/s11098-021-01673-6

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3), 24. https://doi.org/10.1167/9.3.24

Wedgwood, R. (2002). The Aim of Belief. *Philosophical Perspectives*, *16*, 267–297.

Wedgwood, R. (2007). *The Nature of Normativity*. Clarendon Press.

Weigard, A., Clark, D. A., & Sripada, C. (2021). Cognitive efficiency beats top-down control as a reliable individual difference dimension relevant to self-control. *Cognition*, *215*, 104818. https://doi.org/10.1016/j.cognition.2021.104818

Weigard, A., & Sripada, C. (2021). Task-General Efficiency of Evidence Accumulation as a Computationally Defined Neurocognitive Trait: Implications for Clinical Neuroscience. *Biological Psychiatry Global Open Science*, *1*(1), 5–15. https://doi.org/10.1016/j.bpsgos.2021.02.001

Williams, B. (1973). Deciding to Believe. In B. Williams (Ed.), *Problems of the Self* (pp. 136–151). Cambridge University Press.

Yin, L., Reuter, M., & Weber, B. (2016). Let the man choose what to do: Neural correlates of spontaneous lying and truth-telling. *Brain and Cognition*, *102*, 13–25. https://doi.org/10.1016/j.bandc.2015.11.007

Zimmerman, A. Z. (2018). *Belief: A Pragmatic Picture*. Oxford University Press.