Cognitivism, Motivation, and Dual-Process Approaches to Normative Judgment

Brendan Cline
Grand Valley State University

A central source of support for expressivist accounts of normative discourse is the intimate relationship between normative judgment and motivation. Expressivists argue that normative judgments must be noncognitive, desire-like states in order to be so tightly linked with motivation. Normative statements are then construed as expressions of these noncognitive states. In this paper, I draw on dual-process models in cognitive psychology to respond to this argument. According to my proposal, normative judgments are ordinary beliefs that are typically produced by two kinds of process: intuitive-affective processes and domain-general reasoning. When produced by the first kind of process, motivation and judgment tend to align. When produced by the second kind, motivation and judgment might not align. Since the first kind of process is the most common pathway of normative belief formation, normative judgments are typically accompanied by aligning motivation. This proposal enables cognitivists to explain the intimate link between normative judgment and motivation, thereby removing the major obstacle to interpreting normative statements truth-conditionally.

1. The Motivation Argument for Expressivism

It appears that the surface grammar of normative statements supports a truth-conditional interpretation (Glassen 1959). Locutions such as ‘That is wrong’ and ‘He is bad’ look much like ordinary, factual assertions made in the indicative mood, such as ‘That is blue’ and ‘He is alive’. And, as Geach (1960; 1965) famously pointed out, normative terms appear to function much like ordinary descriptive predicates when embedded in complex sentences. A natural interpretation of these (among other) features of our normative discourse is straightforwardly

Contact: Brendan Cline <clineb@gvsu.edu>
truth-conditional: normative statements are ordinary assertions about the normative status of acts, individuals, or states of affairs. A natural companion claim is cognitivism: normative judgments are ordinary beliefs about the normative status of acts, individuals, or states of affairs.¹

Normative judgment internalism (henceforth ‘internalism’) has proven to be a persistent obstacle to this natural pair of claims. When people judge that they ought to ϕ, they are typically motivated to ϕ, and changes in judgment are usually accompanied by appropriate changes in motivation. There also appears to be something amiss about a person who maintains that she ought to ϕ but is not even somewhat motivated to ϕ. These observations are commonly held to support some version of internalism—the view that there is a necessary connection between normative judgments and motivation. An unrestricted version of internalism claims that necessarily, if an agent judges that she ought to ϕ in context C, then she is at least somewhat motivated to ϕ in C. However, it is more common now for internalists to endorse weaker versions of the view that claim that this necessary connection only obtains in fully rational agents, or in normal conditions, or some other restricted set of circumstances.²

The trouble is that according to the popular Humean account of motivation, ordinary beliefs cannot motivate us to act on their own. Rather, independent noncognitive, desire-like states are the source of our motivation, and means-end beliefs merely guide us to particular actions by identifying which ones might be effective instruments for satisfying our desires. But if beliefs are incapable of moving us to act on their own, then normative beliefs could not move us to act on their own. It thus looks like cognitivists are committed to normative judgment externalism (henceforth ‘externalism’)—the view that any relationship between normative judgment and motivation is a contingent feature of our psychology. Externalists are often thought to be incapable of offering a satisfying explanation of the above observations regarding the intimate link between motivation and normative judgment. It therefore looks like normative judgments must instead be noncognitive, desire-like states in order for them to manifest the kind of tight link with motivation noted above.

According to noncognitivists, it is not fruitful to interpret normative judgments in terms of robust truth-conditions, since unlike ordinary beliefs, the non-

¹. I will use ‘normative statements’ as a neutral term to refer to those speech acts—whatever their nature—that involve sincere utterances of declarative sentences that contain normative terms. ‘Normative judgments’ will be a neutral term referring to those mental states—whatever their nature—that are expressed by normative statements. I contrast truth-conditional views from expressivist views about the meaning of normative statements, and contrast cognitivist views from noncognitivist views about the nature of normative judgments. This useful terminology is adopted from Wedgwood (2007).

². See Björnsson, Strandberg, Francén Olinder, Eriksson, and Borklund (2015b) for a review of the various forms internalism has recently taken.
cognitive states that constitute normative judgments are not robustly truth-apt, and are best understood in terms of their motivational effects. From here, a natural view about the semantics of normative discourse is that the meanings of normative statements are to be explained in terms of the noncognitive mental states they express, in the same way that the meanings of descriptive statements can be explained in terms of the beliefs they express. And since normative statements express noncognitive, desire-like states, they too are not fruitfully interpreted in terms of robust truth-conditions. This is the motivation argument for expressivism.3

In this paper, I draw on recent work in cognitive psychology to develop a cognitivist response to this argument. In particular, I show how a dual-process approach to normative judgment enables cognitivist externalists to account for the intimate relationship between motivation and judgment while accepting that beliefs alone are motivationally inert. According to my proposal, normative judgments are (robustly truth-apt) beliefs that are typically produced by two distinct kinds of cognitive process. On the one hand, normative beliefs can be the product of quick, involuntary, motivationally efficacious, intuitive-affective processes. On the other hand, normative beliefs can be the product of slow, deliberate, explicit conscious reasoning. Since the first kind of process is the default pathway for forming normative beliefs, there is a very strong correlation between believing that one ought to φ and being motivated to φ. The explanation for this correlation, however, has to do with the mechanisms typically involved in the production of normative judgments, rather than the nature of such judgments. This framework enables cognitivist externalists to explain all of the relevant phenomena regarding judgment and motivation, thereby blocking the motivation argument for expressivism.

I begin in Section 2 by characterizing dual-process models in cognitive psychology and suggesting an approach for applying them to the psychology of normative judgment. In Section 3, I illustrate this approach by showing how the dual-process framework can enable cognitivists to develop a compelling version of motivational externalism about dangerousness judgments. In Section 4, I extend this framework to normative judgments and present some supporting evidence.

---

3. The central focus of my discussion will be on “pure” expressivist theories (e.g., Blackburn 1984; 1993; 1998; Gibbard 1990; 2003), although I briefly discuss hybrid theories in Section 6 below. The relationship between the core claims involved in this argument are explored in detail in Smith (1994). Note that not all forms of internalism support noncognitivism equally well. Unrestricted versions of internalism lend the strongest support, although Gibbard and Blackburn have appealed to restricted versions of the view. For example, Gibbard (1993: 318–319) maintains that genuine normative judgments typically motivate us in normal circumstances. Despite this weaker claim, the relative frequency of alignment between normative judgment and motivation still raises a puzzle for cognitivists who accept the Humean theory of motivation, and is easily explained by noncognitivists. If one’s internalism becomes too weakened, however, then it is no longer very useful in defending noncognitivism (e.g., Smith 1994; Tresan 2006).
evidence. In Section 5 I argue that my proposal is explanatorily superior to those offered by noncognitivists, and in Section 6 I argue that this result leaves expressivists in pretty bad shape. In Section 7, I defend the a posteriori methodology employed in my argument. The result is a novel approach to cognitivist externalism that overturns the common wisdom regarding the view’s explanatory shortcomings.

2. Dual-Process Models and Normative Judgment

Dual-process models in psychology postulate two distinct kinds of cognitive process in order to explain various psychological phenomena (e.g., Evans & Frankish 2009; Kahneman 2011; Sherman, Gawronski, & Trope 2014). While details vary between different accounts, it is common to distinguish between processes which are (1) quick, subconscious, involuntary, effortless, and independent of working memory, and those which are (2) slow, conscious, deliberate, effortful, and dependent on working memory. The labels ‘Type 1’ and ‘Type 2’ are commonly used to mark the distinction between the first and second kinds of process.5

A key feature of Type 1 processes is that their operations typically take place below the threshold of awareness. It is the results of Type 1 processes, rather than the steps taken within them, that are available to consciousness. These processes are thought to be responsible for a wide range of thoughts we might have, such as recognizing that a sentence is ungrammatical, having a good feeling about a particular stock, recognizing a familiar face, or realizing that a conversational partner has told a joke, to mention just a few. In many cases, thoughts on these matters simply “spring to mind.” And while we often trust these thoughts, we cannot easily explain how we arrived at them—we “just see” the answer. Typi-

---

4. For our purposes, it is not significant that the word ‘cognitive’ is used to characterize these different processes. In this psychological context, ‘cognitive’ roughly means ‘mental’. That is orthogonal to the philosophical distinction between cognitive (i.e., belief-like) and noncognitive (i.e., desire-like) mental states. This is simply an unfortunate consequence of the terminology in which this debate has been couched. Thanks to Tongli Wu for pressing me to clarify this matter.

5. Sometimes, further claims are made about integrated systems composed of Type 1 and Type 2 processes. Here, we can set aside issues surrounding such claims—I intend my proposal to be compatible with, but not dependent upon, (at least some of) these further claims. I am in agreement with Evans and Stanovitch (2013) that the core feature of Type 1 processes is their automaticity, and the core feature of Type 2 processes is their reliance on working memory. The other traits associated with different types of processing are probably best seen as common correlates, rather than necessary criteria. Although my characterization of Type 1 and Type 2 processes diverges in some ways from the work of others (e.g., Mercier & Sperber 2009; 2011), I believe that for the most part, it would be relatively easy to translate my proposal into forms that are more amenable to alternative interpretations of the two kinds of processes.
cally, we do not experience ourselves as voluntarily exerting effort or exercising agency when arriving at conclusions in this way. Rather, the processes that throw up these conclusions often operate involuntarily and automatically, in something like the way that our pupils involuntarily and automatically constrict when a flashlight is pointed at them and dilate when it is moved away. This is not to say that we have no control over such processes, of course. For example, we can direct attention in various ways to “let them loose” on particular problems, in a way similar to how we can voluntarily move a flashlight around to control the patterns of constriction in a healthy eye.

In contrast, the operations of Type 2 processes are generally available to conscious awareness, and require the exertion of effort in order to unfold. The disparity is easy to illustrate with arithmetical problems (Kahneman 2011: 20–21). Consider the problem: $2 + 1 = \_\_$. It is very likely that you can immediately “just see” the solution. Now consider a different problem: $53 \times 34 = \_\_$. If you are able to find the solution to this, it will most likely not involve immediately “just seeing” the solution. Rather, you will need to exert effort in arranging strings of simpler problems, holding the result of each sub-problem in working memory, and then integrating them into a solution. Indeed, perhaps you were feeling averse to investing the effort, and did not even bother to look for the solution. But if you did, you would be aware of each step along the way, and could likely recount each of the steps if you were asked to. It would feel like you were exercising volition in planning how to approach the problem, and while executing the plan.

Type 2 processes are not restricted to arithmetic. Indeed, all of the Type 1 examples mentioned above can have Type 2 counterparts. For example, although we usually recognize our own face immediately upon looking at our reflection, a patient with damage to his fusiform gyrus does not have this luxury, and might instead need to explicitly think to himself, ‘There is a face in that mirror, and no one else is in the room with me, so that must be my face in the mirror.’ It is typically thought that Type 2 processes are more flexible, and largely involve domain-general mechanisms that maintain and manipulate different kinds of information in working memory.6 However, people generally seem to be averse to investing the effort associated with these processes and prefer to avoid doing so. It is therefore thought that Type 2 processes are primarily recruited when the effort is expected to pay off, for example in cases where Type 1 processes issue conflicting results, or when the stakes are high (Braver & Botvinick 2015; Shenhav, Botvinick, & Cohen 2013). As a result, dual-process models often treat the rapid, continuous, specialized activity of Type 1 processes as the “default” pathway, and Type 2 processes as exercising oversight over Type 1 processes.

---

and occasionally intervening when needed (e.g., Evans & Stanovich 2013; Kahneman 2011).

The speed and effortlessness of Type 1 processes is due in part to their domain-specificity, and specialized Type 1 processes across different domains have distinct characteristics that depend on their implementation and proprietary content. For example, computations performed by the mechanisms involved in facial recognition are clearly distinct from those performed by the mechanisms involved in language comprehension. This means that dual-process models of different domains will fill in some of the relevant details in different ways. For example, recent applications of the dual-process framework to the psychology of moral judgment suggest that the mechanisms involved in the Type 1 pathway are implemented by circuitry that is involved in motivation and affective arousal (e.g., Cushman, Young, & Greene 2010; Greene 2014a; Haidt 2001). Imaging studies have associated structures such as the ventromedial prefrontal cortex (vmPFC) and the amygdala with certain patterns of moral judgment characteristic of Type 1 processes (Greene, Nystrom, & Engell 2004; Greene, Sommerville, Nystom, & Cohen 2001; Shenhav & Greene 2014), and damage to these structures has been found to disrupt this pattern (Ciaramelli, Muccioli, Ladavas, & di Pellegrino 2007; Koenigs et al. 2007). The mechanisms involved in Type 2 moral thought, on the other hand, are not similarly linked to affect and motivation, but rather appear to be constituted by domain-general control mechanisms. And indeed, imaging studies have associated structures involved in working memory and cognitive control—such as the dorsolateral prefrontal cortex (dLPFC) and anterior cingulate cortex (ACC)—with distinct patterns of moral judgment, and cognitive load tasks that disrupt working memory disrupt this pattern (Greene, Morelli, Lowenberg, Nystrom, & Cohen 2008; see also Paxton, Ungar, & Greene 2012).

Philosophers have recently devoted much attention to psychologists’ application of the dual-process framework to moral judgment. However, it is less often noted that this approach was largely inspired by the success of the framework in explaining ordinary evaluative thought and decision-making more generally (see, e.g., Bargh & Chartrand 1999; Haidt 2001: 815; Zajonc 1980). And indeed, it now appears that many of the core neural mechanisms implicated in moral thought are largely the same as those implicated in general evaluative thought and decision-making (Fehr & Krajbich 2014; Greene 2014b; Shenhav & Greene 2010). And so while scientists studying evaluative thought and decision making from a dual-process perspective do not always explicitly study things like ought judgments, it seems very plausible that the framework nevertheless is appropriately applied to normative thought in general. In what follows, I take up this assumption in my discussion of the link between motivation and normative judgment, although I acknowledge that it will be important for future
research to more explicitly engage with, for example, thoughts about what one ought to do, in order for the bearing of this framework on the philosophical questions addressed here to be made secure.7

Before moving on, I want to briefly qualify my discussion below. First, it is important to note that my proposal is fully compatible with Type 1 processes being capable of learning based on experience (Railton 2014; Sauer 2012), being rationally unproblematic (Horgan & Timmons 2007; Wielenberg 2014: Chapter 3), and continuously interacting with Type 2 processes (Van Bavel, Xiao, & Cunningham 2012). More importantly, I will aim to maintain a relatively high degree of generality in the basic proposal that I develop. I intend my proposal to be something like an explanation schema that is compatible with a number of distinct ways in which one might fill in the relevant details about the nature of desire, the relationship between emotion and motivation, and so on. In Section 4, I suggest one hypothesis about how aspects of my proposal might be implemented in human brains, and I present some supporting evidence. But it is important to keep in mind that my core proposal is not contingent on that particular hypothesis.

With these qualifications in place, I turn in the next section to illustrate my general proposal by examining the link between judgment and motivation in the domain of dangerousness judgments.

3. Dangerousness Judgments and Motivation

I take it that the dangerousness of an entity or activity is an ordinary, “prosaic” fact that can be more or less correctly or incorrectly represented by ordinary beliefs.8 The semantics of utterances such as ‘That is dangerous’ fit the standard truth-conditional model, and ascribe the property of dangerousness to the refer-

---

7. I do not think this assumption is unfairly hostile toward the views I am criticizing. Since noncognitivists insist on the intimate link between motivation and normative judgment, it seems like they will be inclined to agree that normative judgments are at the heart of what is studied by the science of evaluation and decision making. The inexactness of the explicit questions asked by researchers would not render their work irrelevant to the psychology of normative judgment.

8. A plausible account of dangerousness might be something along the lines of having a disposition to cause harm. McDowell (1988: 175–177) seems to think of the property of dangerousness in normative terms, as the property of meriting fear responses. However, whether or not the dangerousness of a thing merits a fear response is a further fact “above and beyond” the thing’s dangerousness. Fear is not always an appropriate response to danger, and the dangerousness of sharp things does not depend on our reasons to fear them. I therefore think that dangerousness is best thought of as an ordinary, “prosaic” fact. However, even if one prefers to think of dangerousness in normative terms, dangerousness judgments and statements at least appear on the surface to be ordinary beliefs and assertions, and the discussion below illustrates how the motivation that accompanies these judgments can be accounted for within a cognitivist, Humean framework.
ent of the statement. How do we form beliefs about which entities or activities are dangerous? The dual-process framework offers a plausible account (Slovic, Peters, Finucane, & Macgregor 2005).

In some cases, Type 2 processes are clearly responsible. Organizations such as the Center for Disease Control and Prevention, for example, regularly engage in sophisticated data gathering and statistical analysis to determine which entities and activities are dangerous to the public, and just how dangerous they are. Similarly, it is possible for an individual to form dangerousness judgments on the basis of carefully weighing the evidence contained in the reports issued by such organizations. In many situations, however, dangerousness beliefs are produced by intuitive-affective Type 1 processes. A young child, for example, might come to believe that hot stoves are dangerous immediately after burning her hand on one. Along similar lines, a thing’s affective salience can influence people’s beliefs about how dangerous it is. In cases like these, a motivationally efficacious desire-like state (e.g., pain, fear) precedes and directly contributes to the formation of the dangerousness belief.

According to this approach, then, the judgment that something is dangerous is a belief which can be produced by two distinct kinds of process. Once such beliefs are formed, they are made available to working memory and can be involved in reasoning in the same way that other ordinary beliefs can. The semantics of dangerousness statements is to be treated in the ordinary truth-conditional manner: typical dangerousness statements are assertions about the way the world is, and ‘dangerous’ retains a univocal meaning when embedded in complex sentences.

If we are Humeans about motivation, however, then this approach will commit us to the externalist claim that a belief that \( x \) is dangerous is insufficient to generate motivation regarding actions related to \( x \). And this might seem odd, since dangerousness judgments do seem relevant to the choices that people make. People are typically at least somewhat motivated to avoid things they regard as dangerous, and it would be an odd character who systematically neglect-

---

9. I believe that assertions containing the word ‘dangerous’ can be used to express one’s fear, or to instruct one’s interlocutor to behave in certain ways. But these cases can usefully be interpreted along pragmatic lines, in the same way that asserting that an object is expensive can serve to express one’s concern about the object in question, and direct one’s audience to treat the object with care.

10. I do not maintain that it is impossible for other Type 1 processes to result in the tokening of a dangerousness belief. For example, long-term memory recall of a dangerousness belief might involve Type 1 processes, and the recalled judgment might have initially been formed via Type 2 processes. To help clarify, I will use the phrase ‘intuitive-affective’ to pick out the subset of Type 1 processes where motivationally efficacious states are involved in giving rise to the conscious thoughts that are the outputs of the process. I will sometimes use this qualifier to draw attention to this precisification. But unless otherwise noted, when I discuss the Type 1 pathway of normative judgment formation below, I will be referring to intuitive-affective Type 1 processes.
ed all of her dangerousness judgments in her decision-making. Is this a serious problem with claiming that dangerousness judgments are beliefs? Might it be better to instead consider pursuing an expressivist semantics for dangerousness statements? I think not. The dual-process framework enables us to understand the motivation associated with dangerousness judgments from a cognitivist perspective, even within the constraints of the Humean account of motivation.

If beliefs about what is and is not dangerous are typically generated or activated by motivationally efficacious, intuitive-affective Type 1 processes, then we can give a cognitivist explanation of why specific motivations tend to accompany beliefs about whether or not a thing is dangerous. If my belief that \( x \) is dangerous was produced or activated by intuitive-affective Type 1 processes, then I will be motivated to avoid \( x \) or proceed with caution when dealing with \( x \). This is because motivationally efficacious states were involved in the production of this belief. And since intuitive-affective Type 1 pathways are the most common, default processes involved in the generation of beliefs about dangerousness, people are typically motivated to act in ways that align with their beliefs about how dangerous things are. (This motivation will be modulated by the individual’s level of risk aversion, of course.)

At the same time, it would be implausibly strong to claim that people are always (even at least somewhat) motivated to act in accordance with their judgments about what is and is not dangerous. For example, some people are terrified of flying in planes and are unwilling to fly, despite sincerely believing that flying in planes is not very dangerous. In the other direction, some people are uninhibited about regularly engaging in activities—such as smoking or driving a car—which they know to be dangerous. In the other direction, some people are uninhibited about regularly engaging in activities—such as smoking or driving a car—which they know to be dangerous. Any plausible account of the psychology of dangerousness judgments needs to be able to account for such phenomena. Fortunately, the dual-process framework offers us the resources to do this.

When the formation of a dangerousness judgment does not engage intuitive-affective Type 1 processes, then one can form the belief that \( x \) is (is not) dangerous without having any corresponding motivation appropriate to \( x \)’s dangerousness (innocuousness). Indeed, what is common to the examples

---

11. There are nearby phenomena that are relevant as well. If I believe that something is likely to cause me to die, and I desire to continue living, then I will be motivated to avoid that thing. Plausibly, it is precisely such desires that commonly give rise to our dangerousness judgments. However, these kinds of cases pose no puzzle for the Humean about motivation. The problem cases are those where certain classes of judgment appear to be both belief-like and sufficient for motivation. This is why dangerousness judgments pose an interesting puzzle for motivational Humeans. Thanks to James Beebe for pressing me on this point.

12. Many of these kinds of cases will simply involve conflicting motivation, where some pro tanto motivation to fly or avoid smoking is simply overridden by stronger competing desires. However, it is also plausible that at least on occasion, an agent makes genuine dangerousness judgments despite completely lacking any accompanying motivation.
of judgment-motivation mismatch that I just described is that they are usually cases where people reason themselves (via Type 2 processes) to conclusions regarding the dangerousness of something, but fail to engage relevant Type 1 processes. Someone who reads text warnings about the dangers of smoking, for example, may be less inhibited about it than someone who is repeatedly exposed to graphic images of smoking’s ill effects (Borland et al. 2009). Similarly, reasoning about the comparative safety of flying and driving will often do little to help a person who is truly afraid of flying but not driving. However, if you put such a person in an empty plane on the ground, and rock the plane to associate turbulence with a safe environment, you may have some success in gradually changing their willingness to fly (Hajek 2016).

This framework does not entail that it is impossible for Type 2 processes to influence motivation, of course. While I agree with the Humean thesis that belief alone is motivationally inert, it is possible to voluntarily re-frame a situation, or direct attention in such a way that the operations of Type 1 processes change one’s motivation. For example, you can intentionally subject yourself to images of people who are visibly suffering from the effects of smoking-induced disease. Similarly, if you are experiencing terror at 10,000 feet, you might not be able to shut it down voluntarily. But you can close your eyes and picture yourself somewhere pleasant, or try to imagine that you are actually on the ground and people are just rocking the plane—offering new input to involuntary Type 1 process. These kinds of interventions can act in a manner that is analogous to moving a flashlight around with your hands to control your pupil’s involuntary constriction and dilation.

A person’s belief about a thing’s dangerousness, then, is insufficient to determine their motivation. What ultimately determines their motivation is the source of that belief. If it is produced by intuitive-affective Type 1 processes, it is likely that motivation and judgment will match. If it is produced largely by Type 2 reasoning processes without the engagement of intuitive-affective Type 1 processes, it is possible that motivation and judgment will not align. The important point to keep in mind is that it is in the processes that are responsible for the formation of the judgment, rather than in the judgment itself, that the story about motivation is to be found.

4. Extending the Account to Normative Judgment

I maintain that a story similar to the one told above about dangerousness judgments can be told about normative judgments. The first thing to notice is that if correct, the sketch of dangerousness judgments in the previous section shows us that there is a mechanism that transforms intuitive-affective responses into
beliefs about things in the world, and the products of this mechanism are made available to working memory for domain-general reasoning. If we accept that story, then we are already committed to such a thing going on in our minds at least in some cases. My proposal now is that a similar mechanism is involved in many of our normative judgments as well.

How might such a mechanism be implemented in the brain? Recent research has generated a number of suggestive results. It now appears that signals broadcast from (a) circuitry involved in reward learning and action selection (e.g., the basal ganglia, midbrain dopaminergic nuclei) and (b) circuitry involved in emotional arousal and motivation (e.g., the amygdala, hypothalamus, anterior insular cortex) to (c) the vmPFC are key to understanding the intuitive-affective Type 1 pathway of normative judgment formation (for details, see Camerer, & Montague 2008; Craig 2009; Haber & Knutson 2010; Hare, Camerer, Knoepfle, O’Doherty, & Rangel 2010; Rangel & Clithero 2014; Rangel, Railton 2014; Sanfey, Rilling, Aronson, Nystrom, & Cohen 2003; Smith et al. 2010). It seems that once this information reaches (c) the vmPFC, it is integrated into explicit normative thoughts that can become available to conscious awareness and working memory, and can be verbally expressed by normative statements (see Basten, Biele, Heekeren, & Fiebach 2010; Roy, Shohamy, & Wager 2012; Shenhav & Greene 2014). There is reason to think that our desire-like, motivationally efficacious states are realized by (a) and (b) circuitry (Arpaly & Schroeder 2014: Chapter 6; Schroeder 2004; Schroeder, Roskies, & Nichols 2010). Consistent with this view, imaging research suggests that these structures continuously and automatically appraise actual and potential actions and states of affairs (Lebreton, Jorge, Michel, Thirion, & Pessiglione 2009), and that future behavior can be predicted on the basis of automated activity in these structures (Abe & Greene 2014; Haruno & Frith 2010; Tusche, Bode, & Haynes 2010). These (a) and (b) structures thus seem to simultaneously (i) regulate affect and guide action selection, and (ii) generate consciously accessible normative thoughts. My suggestion is that it is the fact that mental states realized in (a) and (b) circuits are the common cause of (i) and (ii) that explains the intimate connection between normative belief and motivation.

At the same time, research has found that evaluations and choices are sometimes influenced by (d) structures involved in working memory and cognitive control (such as the dlPFC and ACC). It appears that these circuits constitute distinct control systems that can compete with or influence the activity of the (a)–(c) circuits. For example, purposefully focusing one’s attention on the healthiness of food options can affect one’s eating choices, and this effect appears to be mediated by activity in the dlPFC modulating activity in the vmPFC (Hare, Malmaud, & Rangel 2011; see also Baumgartner, Knoch, Hotz, Eisenegger, & Fehr 2011; Botvinick & Braver 2015; Hare et al. 2009; Hutcherson, Plassmann, Gross,
Rangel 2012; Ocshner & Gross 2005; Plassmann, O'Doherty, Shiv, & Rangel 2008; Rudorf & Hare 2014; Shenhav et al. 2013; Sokol-Hessner, Hutcherson, Hare, & Rangel 2012). These changes can also be broadcast back to (a) and (b) circuitry, thereby influencing one’s motivation (Pujara & Koenigs 2014). This kind of effect is analogous to my earlier examples of intentionally pointing a flashlight at one’s eye to control pupil constriction, or intentionally imagining a pleasant, safe place to mitigate one’s fear response while flying. My preferred interpretation of these “flashlight interventions” is that effortful activity of Type 2 processes offers new inputs for appraisal by involuntary intuitive-affective Type 1 processes.

However, although it is possible for effortful activation of (d) circuitry to influence motivation and behavior, it appears that these effects are indirect, and that the mechanisms involved in conscious reasoning and working memory are dissociable from those more directly involved in the regulation of motivation and affect. In the first place, structures involved in domain-general working memory—such as the dIPFC—do not seem to have any special relationship to motivation, as they are involved in manipulating information in working memory across a number of domains (Barbey et al. 2013; Botvinick & Braver 2015; Chein et al. 2011; Fedorenko et al. 2013). Moreover, whereas studies of evaluative thought regularly implicate (a)–(c) circuitry, (d) circuits are only occasionally found to be activated. This suggests that Type 2 processing is not necessary for the formation of evaluative judgments and decisions, but rather an optional extra layer that can be recruited when needed (Haber & Knutson 2010).

Many now accept that sometimes, people fail to be motivated to act in accordance with their normative judgments. And I suggest that Type 2 processes implemented in the dIPFC might be able to account for these cases as well. Via explicit reasoning, an agent might come to a conclusion about what she ought to do, and yet this might be motivationally epiphenomenal, since activity in the dIPFC might not interact with or influence circuits involved in motivation and affective arousal.

Evidence for the existence of this motivationally epiphenomenal Type 2 activity comes from studies of patients with vmPFC damage. One landmark study observed a group of vmPFC patients that manifested an inability to integrate their explicit thoughts about optimal decision strategies in a gambling task with their performance on that task (Bechara, Damasio, Tranel, & Damasio 1997). This was in contrast to healthy subjects, who were capable of learning and implementing optimal strategies before becoming able to explicitly articulate the strategy or explain why it was optimal. Indeed, optimal strategies were employed even by healthy subjects who never became able to explicitly state the merits of the strategy. It appears that healthy subjects could just “immediately see” the optimal strategy via intuitive-affective Type 1 processes, and this explains why their motivation matched their judgments about which strategy was best. On
the other hand, it seems that the vmPFC patients may have explicitly reasoned about the task, formed a belief about which strategy was best, and yet continued to pursue the suboptimal strategy because their explicit (Type 2) reasoning had not engaged the motivational circuitry underlying the Type 1 pathway.

Subsequent studies of vmPFC patients have continued to document notable abnormalities in behavior in experimental games (e.g., Krajibich, Adolphs, Tranel, & Denburg 2009; Koenigs & Tranel 2007) and responses to affectively charged scenarios (Koenigs et al. 2007). However, probes of vmPFC patients’ explicit attitudes about abstract, hypothetical choice scenarios have found considerable overlap with healthy controls (Leland & Grafman 2005). Additionally, vmPFC patients are often indistinguishable from healthy subjects in a range of measures of intelligence and general cognitive ability, suggesting that declarative memory and working memory are preserved (Damasio 1994). The problem with vmPFC patients, it seems, is not that they are unable to emote and perform motivated actions, nor that they are unable to explicitly reason about practical matters, but rather that these two aspects of their psychology cannot interact fluidly. The damage to the vmPFC seems to disrupt the connection between explicit conscious reasoning and motivation. Considerations such as these support the interpretation of the vmPFC as a way-station for relaying information between the Type 1 processes involved in motivation and affect and the Type 2 processes involved in normative reasoning (see Roy et al. 2012).13

I now want to step back and bring this discussion into conversation with the philosophical questions with which we began. I propose that an interpretation along the lines suggested in the previous section on dangerousness judgments is apt. According to this approach, normative judgments are ordinary beliefs that are often tightly linked with motivation, but not essentially so. The key to understanding the relationship between normative judgment and motivation is in the processes responsible for the production of the judgment, rather than in the nature of the judgment itself. For normative judgments that are produced by intuitive-affective Type 1 processes, motivation and judgment will typically align. The explanation for this is that the judgment is produced by motivationally efficacious, desire-like aspects of our psychology. Since this is the default process of normative belief formation, people are usually motivated to act in accordance with their normative judgments.

As with dangerousness judgments, however, it would be implausible to maintain that normative judgments are always accompanied by (even some) corresponding motivation. And again, the current framework offers a natural ex-

13. Note that I am not claiming that vmPFC patients are counterexamples to internalism (cf. Roskies 2003). Rather, I am using these cases to support the claim that the ordinary mechanisms underlying Type 2 normative thought are dissociable from those underlying Type 1 normative thought, and these mechanisms have different motivational profiles.
planation for cases of judgment-motivation mismatch. In particular, normative judgments that are produced by explicit, conscious, Type 2 reasoning might not align with motivation. The explanation for this is that beliefs alone are motivationally inert, and Type 2 processes are not implemented by circuits directly involved in the generation of affect and motivation. When reasoning and deliberation does influence motivation, this is not because beliefs about what one ought to do have a special motivational force as part of their essential nature. Rather, it occurs due to the ways in which the various contents of our conscious thoughts (such as the healthiness of our food options) interact with our desires and affective dispositions when the contents of those thoughts are subjected to intuitive-affective Type 1 appraisal.14

The important point for our purposes is the abstract level of explanation given in the previous two paragraphs. The core claim is that motivationally efficacious, desire-like states are the typical source of normative beliefs, and this is the basic explanation for the tight link between motivation and judgment. If something relevantly like this explanation-schema works, then cognitivists can explain the intimate link between motivation and normative judgment within a Humean framework. And at present, it seems that there is relatively good evidence for this general idea.15

5. Do Noncognitivists Do Better?

As we saw in Section 1, the motivation argument for expressivism rests on the claim that, in light of the Humean account of motivation, only noncognitivists

14. The main focus of this discussion is on first-personal judgments of the form ‘I ought to φ’. The reason for this is that these are the judgments that typically come under the purview of internalist theses. However dual-process models are thought to be applicable to all sorts of evaluative thoughts one might have, such as the quick impressions we form about people upon first meeting them (Zajonc 1980). Usually, appraisals and associated motivations—e.g., whether to trust someone—occur rapidly and without the need for explicit reasoning, although it is possible to explicitly reason about things such as whether or not to trust a new acquaintance.

15. Recently, Sinhababu (2017: Chapter 4) has independently developed a similar approach to defending cognitivist externalism, and I take my work to largely complement his by integrating dual-process models in cognitive psychology, and offering a detailed hypothesis regarding the neural implementation of the relevant processes. Aside from his focusing mainly on moral judgments (rather than normative judgments generally), the major difference between our accounts is in the details of what I am describing as the ‘intuitive-affective’ pathway. Essentially, Sinhababu treats emotional/affective responses themselves as the source of intuitive-affective moral beliefs, whereas my proposal treats activity in structures involved in emotion—(b) circuitry—as common causes of physiological changes and normative beliefs. I also propose that activity in structures involved in reward and action selection—(a) circuitry—plays an important role in directly causing normative belief formation, and that felt emotions or affective states are not necessary for this to occur. Thanks to Neil Sinhababu for helping me to get clearer on these details.
can adequately explain the intimate link between normative judgment and motivation. In the previous section, I have developed a cognitivist explanation of this intimate link within the constraints of the Humean framework. I take it that this proposal is sufficient to blunt some of the force of the motivation argument. However, if noncognitivism still offers the best explanation of the relevant phenomena, this would still be a significant shortcoming of cognitivism. In this section, I argue that my cognitivist proposal in fact offers a better explanation of the relevant phenomena than noncognitivists do.

To begin, note that there are two basic patterns of motivation that an account of normative judgment needs to explain. The first is that judgment and motivation tend to align in most people most of the time. Call this the alignment pattern. The second is that this alignment does not always obtain every time in every agent. Call this the mismatch pattern. It seems that most parties to the debate accept this much.\(^{16}\)

According to noncognitivists, to judge that one ought to \(\phi\) just is to be in a desire-like mental state that motivates one to \(\phi\). So, if a person judges that she ought to \(\phi\), she is motivated to \(\phi\). Moreover, changes in judgment amount to changes in motivationally efficacious desire-like mental states, and this explains why they tend to co-vary. This is a strong explanation of the alignment pattern.

According to my cognitivist proposal, to judge that one ought to \(\phi\) is to be in a motivationally inert state of belief with the content that one ought to \(\phi\). However, the standard cause of normative beliefs are intuitive-affective Type 1 processes that simultaneously motivate one to \(\phi\) and cause one to form the belief that one ought to \(\phi\). Furthermore, when the belief that one ought to \(\phi\) is produced by Type 2 processes, this often involves tokening affectively salient representations that serve as inputs to intuitive-affective Type 1 processes, which influence the agent’s motivation. The result is the alignment pattern.

It therefore seems that both accounts can adequately capture the alignment pattern. What about the mismatch pattern?

My cognitivist proposal offers a straightforward explanation of this pattern as well. For it allows that normative beliefs are sometimes produced by Type 2 processes that result in the formation of a normative belief without engaging mechanisms directly involved in affect and motivation. And since beliefs alone are motivationally inert, merely forming a normative belief is insufficient to bring about a corresponding motivation. Hence the mismatch pattern.

Noncognitivists, in contrast, do not have as easy a time explaining cases where normative judgment and motivation do not align, because their account

\(^{16}\) In particular, it seems that these patterns are accepted by both Blackburn (1998: 60ff) and Gibbard (1993: 318–319; 2003: 153–154; see also Strandberg 2012), even though some authors (e.g., Bromwich 2013) have recently advocated for unrestricted versions of internalism. The target of my argument in this section will be noncognitivists who accept restricted versions of internalism.
treats normative judgments as a kind of motivationally efficacious state. For example, Blackburn (1998: 59–68; see also Bedke 2009; Eriksson 2014) suggests that although individuals might sometimes fail to be motivated by their normative judgments, those individuals could only make genuine normative judgments by participating in a community in which normative judgments typically align with motivation, or by possessing a psychology in which normative judgments typically do align with motivation. It is unclear that these explanations are in line with the commitments of noncognitivism, however. For they seem to allow that a person can make a normative judgment despite failing to be in a relevant desire-like, motivationally efficacious state. But if normative judgments just are desire-like, motivationally efficacious states, then it seems that these individuals must instead be in some other kind of state. In that case, they would not actually be manifesting the mismatch pattern. So, this does not really seem to be a noncognitivist explanation of the mismatch pattern at all.

In light of these kinds of concerns, Gibbard (1993: 318–319; 2003: 152–158) has proposed that the mismatch pattern arises when the motivationally efficacious states that constitute normative judgments become somehow distorted, such that their motivational effects are blocked (see also Björnsson & McPherson 2014; Eriksson 2014; Toppinen 2015; cf. Ridge 2015; Strandberg 2012). So, while the agent’s judgment is the kind of state that is disposed to have motivational effects in normal conditions, those dispositions merely get suppressed in some abnormal cases. Hence the mismatch pattern.

Even though this could potentially explain the mismatch pattern, it is not a particularly strong explanation. First, notice that my cognitivist proposal maintains that the mismatch pattern arises only in less common cases where belief-like mental states that are typically accompanied by matching motivation fail to be so accompanied. This hypothesis is supported by the fact that the pattern consists in cases where there is judgment but not motivation—the lack of motivation is evidence of a lack of desire. Now consider: what is the evidence for supposing that the mental state an individual is in when he manifests the mismatch pattern is a blunted noncognitive, desire-like state? The motivational profile of the agent certainly does not support this interpretation. Instead, it seems that noncognitivists can only appeal to their ability to explain the alignment pattern, and then claim on inductive grounds that all normative judgments must be noncognitive, desire-like states. However, the alignment pattern does not necessarily tell in favor of noncognitivism over my cognitivist proposal, since that pattern is straightforwardly explicable within both frameworks. And there does not seem to be any independent evidence for positing blunted desire-like states in individuals manifesting the mismatch pattern. As a result, this noncognitivist explanation of the mismatch pattern seems to merely accommodate the mismatch pattern, rather than be supported by it. In contrast, my cognitivist proposal is
supported by the mismatch pattern, and can cleanly explain the pattern in terms of the nature of normative judgment and the operation of the motivationally epiphenomenal activity of Type 2 processes.

I conclude that cognitivists are in a better position than noncognitivists to explain the relevant phenomena. Armed with the dual-process model of normative judgment, cognitivists can offer a precise and informative account of both the alignment pattern and the mismatch pattern. On the other hand, while noncognitivists can give straightforward explanations of the alignment pattern, they must backpedal to accommodate the mismatch pattern. As a result, observations about the relationship between judgment and motivation do not constitute an obstacle for truth-conditional interpretations of normative discourse, because cognitivists are in fact best placed to explain these observations. As I argue in the next section, this result leaves expressivism in pretty bad shape.

6. Consequences of the Account

In the previous section, I argued that my cognitivist proposal offers a better explanation of the mismatch pattern than noncognitivism does. To see the full import of my proposal, however, let us briefly suppose that my case against the noncognitivists’ explanation of the mismatch pattern was not convincing. Suppose instead that the two accounts are roughly equally matched in terms of explanatory ability.

Even in this case, things do not look good for the noncognitivism-expressivism package. The basic trouble is that the expressivist semantic theory faces some very serious problems. As I observed in Section 1, normative language appears, on the surface, much like ordinary descriptive language. The surface grammar of normative statements looks very much as if they are assertions about the normative status of various acts, individuals, or states of affairs. Furthermore, normative statements appear to retain a univocal meaning when embedded in complex sentences (Geach 1960; 1965). And despite some creative responses to the Frege-Geach problem, expressivists have failed to elaborate an adequate treatment of the meanings of complex normative sentences, and have failed to adequately explain the apparent continuity between normative and descriptive statements. Indeed, there is some reason to think that no expressivist account can be made to work as a semantic model of natural languages like English (Schroeder 2008). At the very least, pursuing an expressivist semantics involves taking on significant theoretical burdens that are not shared by cognitivists who advocate a truth-conditional interpretation of normative discourse.

In the face of these difficulties, internalism has been instrumental in preserving the appeal of expressivism, since it has appeared that cognitivists sim-
ply cannot account for the pervasive role of motivation in normative thought (e.g., Gibbard 1990: 33; 2003: Chapter 1). However, if the framework proposed in this paper is on the right track, then this charge against cognitivists is mistaken. Even within the constraints of the Humean theory of motivation, cognitivists are able to give a straightforward explanation of the role of motivation in normative thought and discourse. Furthermore, by treating normative judgments as beliefs, cognitivism allows us to specify the meanings of normative statements without incurring the theoretical costs of an expressivist semantics.

This result leaves expressivism in pretty bad shape. At best, cognitivists and noncognitivists are on par in terms of the explanations they can offer for the link between judgment and motivation. At worst, cognitivists can offer better explanations. Thus, one of the key motivations for pursuing an expressivist semantics is removed. But given that expressivism is a problematic account of the meaning of normative statements, it is now much less clear why we would find such accounts attractive. For any of the other purported benefits of expressivism—such as legitimizing our normative discourse in a world without irreducibly normative facts—can be offered by a number of alternative views that accept truth-conditional accounts of normative discourse (e.g., Joyce 2001: Chapter 8; Olson 2014: Chapter 9; Railton 1986; Schroeder 2007). Indeed, even the coordinating functions of normative discourse can readily be explained without incurring the costs of an expressivist semantics (Bar-On & Chrisman 2009; Strandberg 2011). I conclude that the account of normative judgment proposed above represents a significant threat to the viability of the expressivist program.

It might seem that hybrid semantic views that attempt to combine truth-conditional and expressivist elements will fare better, since they are not necessarily committed to the problematic semantics of “pure” expressivism (e.g., Joyce 2006: Chapter 2; Ridge 2015). However, I believe that the appeal of these views is also compromised by the proposal developed here. Typically, the purpose of building in the truth-conditional element is to avoid the semantic problems faced by expressivists, and the purpose of building in the expressivist element is to retain the explanatory power of noncognitivism on the topic of motivation. But if cognitivists can explain all of the motivational effects that need explaining, then there no longer seems much point in retaining the unnecessary machinery associated with the expressivist element.

Similar considerations apply to hybrid accounts of normative judgment which attempt to blend elements of cognitivism and noncognitivism (Kumar 2016; see also Kauppinen 2015; Kriegel 2012). Kumar (2016), for example, defends the view that moral judgments are a natural kind composed of both belief-like and desire-like mental states, arguing that this approach promises to capture the benefits of both cognitivism and noncognitivism. However, when compared to the proposal I have developed here, hybrid accounts of judgment such as
Kumar’s appear to be in the same boat as hybrid semantic accounts in that they seem to build in unneeded machinery. If we accept at least the cognitivist and truth-conditional component of these views, then it is unclear what additional benefits are provided by adding in a noncognitivist element, given that my proposal can handle the relevant motivational effects equally well. Thus, the proposal developed here is threatening to the viability of hybrid views of normative judgment as well.

Before moving on, I would like to briefly clarify the role that the dual-process framework is supposed to play in my argument. I do not claim that cognitivism follows directly from dual-process approaches to normative judgment. Nor do I think that dual-process models are necessarily incompatible with noncognitivist or hybrid accounts of normative judgment. Instead, I have primarily appealed to the dual-process approach in a defensive manner, proposing an interpretation of the relevant empirical models that enables cognitivist externalists to explain the intimate link between normative judgment and motivation. The goal of this main argument is simply to respond to the most serious standing challenge facing cognitivism: explaining the intimate link between judgment and motivation within the Humean theory of motivation. However, what I’ve argued in this section is that when this challenge is adequately responded to, the balance of evidence favors cognitivism, given the independent problems with expressivism. And since hybrid accounts do not have any clear explanatory advantage over my pure cognitivist account, there seems to be no reason to go beyond pure cognitivism and build in an expressivist or noncognitivist component. Thus, when my central defensive argument is considered in the larger dialectical context, cognitivism emerges as the view with the most overall plausibility.17

7. Internalism as a Conceptual Claim

As we saw in Section 1, internalism is a key premise in the motivation argument for expressivism. I have argued that the dual-process approach to normative judgment developed here enables cognitivist externalists to account for all of the relevant phenomena concerning the intimate link between normative judgment and motivation, thereby blocking the conclusion of the motivation argument. However, it might be argued that my approach has no bearing on the debate between internalists and externalists. This is because my primary aim has been to illuminate the cognitive mechanisms underlying normative thought in ordinary humans. But internalism is usually interpreted as a conceptual claim about the nature of normative judgment, and as such is immune to empirical

17. Thanks to an anonymous reviewer for pressing me to clarify this point.
refutation. Internalists do not claim that it just so happens that normative judgments are usually linked with motivation, but rather that in order to count as a normative judgment, a mental state must bear some necessary connection with motivation. And since I have offered no direct arguments that this conceptual claim is false, the motivation argument for expressivism remains untouched (see Kennett & Fine 2008; Smith 2008).  

While it is true that internalism is typically understood as a conceptual claim, I do not think that this line of objection is very productive. My goal has been to sketch an empirically motivated account of what goes on in our heads when we form thoughts of the form ‘I ought to ϕ’. I have argued that this account offers plausible explanations of not only the pervasiveness of the alignment pattern, but also explains the less common mismatch pattern. In response, the internalist is objecting that our concept of a normative judgment excludes my account \textit{a priori}, because my account construes the connection between judgment and motivation as contingent, rather than necessary. No account like mine could possibly be true, because it flouts this internalist conceptual constraint. But what evidence is there for thinking that an account like mine could not possibly be correct?

One line of evidence might be the alignment pattern described above. One might think that only if internalism is true could there be such a pervasive link between judgment and motivation. But this evidence is unpersuasive, since I have already shown how externalists could explain such a pattern. More importantly, it is unclear that those observations would be enough to show that normative judgments must bear some necessary relation to motivation. As Smith (2008: 210) points out, even if every actual instance of judgment were matched with motivation, this would not show that judgment and motivation are linked as a matter of conceptual necessity.

A more promising line of evidence might seem to come from the fact that internalists do not feel comfortable attributing a genuine normative judgment to an agent who says or thinks ‘I ought to ϕ’ and yet remains wholly unmotivated to ϕ. This might be taken to support the idea that our concept of normative judg-

---

18. Thanks to Alex King for pressing me on this point. While some internalists, such as Smith (1994: Chapter 3) and Wedgwood (2007: Chapter 1), maintain that the most plausible version of internalism appeals to irrationality to account for cases where motivation and judgment fail to align, this maneuver does not seem to be available to expressivists. This is because they typically maintain that normative notions have no place in the fundamental analysis of normative thought, and that normative judgment and motivational internalism can be understood without the use of normative terms like ‘rational’. Thus, rationality-based versions of internalism are outside of the scope of this paper, since my main goal here is to defend cognitivism. It is the psychological—not normative—interpretations of internalism that are most relevant here. See Cline (in press) for further discussion of Smith’s “practicality requirement” in the context of dual-process models of moral judgment.
ment does not legitimately apply to mental states that do not bear some necessary connection with motivation.

This evidence, however, is rather weak, since not all parties accept that the concept of normative judgment cannot legitimately be applied in such cases. It is true that sometimes some individuals might not understand certain concepts, and thus might make mistakes in their application. For example, a badly confused person might maintain that cats are a kind of musical instrument. In cases like that, something has clearly gone wrong. However, internalism is a notoriously controversial thesis, and it is far from clear that externalists are confused in the way that a person claiming cats to be musical instruments would be (cf. Francén 2012). Indeed, in a major survey of professional philosophers, roughly 30% of respondents claimed to endorse or to lean toward externalism, around 35% endorsed or leaned toward internalism, and about 35% answered “other” (Bourget & Chalmers 2014: 496). This lack of consensus suggests that one should be especially cautious about pressing charges of conceptual confusion against either side.

Perhaps externalist philosophers’ concepts of normative judgment are theory-laden and have become biased by their theoretical commitments. However, that hypothesis is not well supported by work in experimental philosophy on folk intuitions about moral judgment (e.g., Björnsson, Eriksson, Strandberg, Francén Olinder, & Björklund 2015a; Strandberg & Björklund 2013). It seems that at best, some people sometimes have some hesitancy in attributing genuine moral judgments to individuals who lack corresponding motivation. This is not very helpful to internalists, since there is certainly not uniform refusal to make such attributions under the conditions that have been studied, unlike what we would expect for obvious mistakes like the cats-as-musical-instruments thesis. Furthermore, what hesitancy there is can easily be explained within an externalist framework.

For example, if my proposal is on the right track, then the mismatch pattern will be relatively uncommon. As a result, there are various ways in which general pragmatic mechanisms could explain why expressing one’s moral beliefs usually conveys information about one’s motivations (Strandberg 2011; see also Bar-On & Chrisman 2009). This would explain our expectation that agents’ judgment and motivation align even if there is no necessary connection between motivation and normative judgment.

19. Indeed, most internalists now agree that it can sometimes be legitimate to attribute a normative judgment to an agent who is not motivated to act accordingly, as long as that agent is irrational, or usually motivated to act in accordance with his judgments, or is a member of a community of individuals that are usually motivated to act in accordance with their judgments. However, the profusion of different internalisms seems to call into question the assumption that there is a definitive fact of the matter about the boundaries of a single, crisp concept shared by all parties.
Similarly, we have known for some time that people do not treat every member of a category the same way. Instead, people have an easier time identifying more prototypical instances as members. For example, subjects are quicker in identifying a robin as a kind of bird than in identifying a penguin as a kind of bird (Rosch 1978). Given that most people are usually motivated to act in accordance with their normative judgments, motivation might be to normative judgment as flight is to birds. Hence our greater hesitancy in attribution.

Furthermore, there undoubtedly are cases where people insincerely claim to believe some normative claim. We are on the lookout for people who try to manipulate us (Tooby, Cosmides, & Price 2006), and one way someone might do that is by claiming to share our values. As a result, it makes sense for us to be suspicious of people who are unmoved by what they purport to think is good and right.

In light of considerations such as these, it is not difficult to see why some philosophers and laypeople would have seemingly internalist intuitions, even if externalism were true. I therefore think there is little reason to maintain that my proposal is irrelevant on the grounds that it violates our concept of a normative judgment. Clear-cut conceptual mistakes might be firm ground to argue from. But in more controversial cases, one’s opponent cannot so easily be dismissed. Instead, it seems more productive to draw on empirically-informed accounts about mechanisms underlying normative thought to help shed light on the question of whether normative judgments are constituted by noncognitive states (see Bromwich 2013; Kumar 2015; 2016; Prinz 2015).

Indeed, I believe that possessing a sufficiently detailed account along these lines is a prerequisite for resolution of this debate. This is due to the common claim that the internalist thesis only requires that an agent possess a smidgen of motivation in order for internalism to be satisfied. What is a smidgen of motivation? How could we determine whether or not a smidgen of motivation is present if it never influences action? I cannot imagine how armchair interpretation of imaginary cases could possibly offer the final story on this issue. Instead, it seems that the story must ultimately implicate the operations of some cognitive mechanism or other that plays some role in human action. By focusing on the mechanisms underlying normative thought and identifying their functional profiles, we can make progress beyond just guessing whether or not an imaginary agent had a smidgen of motivation to φ despite not actually φ-ing (cf. Botvinick & Braver 2015: 104–105).

It might be thought that one advantage of accepting internalism as a conceptual claim is the ability to appeal to the semantic content of normative judgments in order to explain why it is that motivation is relevant to normative judgments.20

20. Thanks to an anonymous reviewer for highlighting this potential advantage.
This may indeed be something externalists cannot easily offer (but see Sturgeon 1999: 95). However, by interpreting normative beliefs as typically caused by motivationally efficacious desire-like mental states, the dual-process framework developed here promises to explain why it is that we experience certain potential actions as calling out for responses from us. Since, on my proposal, desires and affective responses cause us to be motivated in various ways, and also cause us to think that we have reasons to act in those ways, there is a readily available explanation for why there is a tight link between normative judgments and motivation. Even if one does not find this explanation sufficiently satisfying, this advantage of accepting internalism as a conceptual claim does not seem to be a decisive consideration in its favor. Views on these matters need to be assessed on their overall plausibility, and this advantage would only be one part of the equation.

Finally, even if internalism is a true conceptual claim, it is possible that it is a conceptual claim in need of revision. It is not uncommon for scientific advances to lead us to refine our concepts in response to new evidence. For example, despite once-popular opinion to the contrary, whales turned out not to be a kind of fish. If we are interested in understanding the actual normative judgments that people make, it seems at least conceivable that a deeper understanding of the underlying cognitive mechanisms could show us that our concepts were mistaken (see Bedke 2009: 202–203; Kumar 2016). Rather than latching on to a real necessary connection, perhaps we merely overgeneralized on the basis of the constant conjunction of judgment and motivation.

The framework developed above enjoys a fair degree of empirical support, and offers informative explanations of several key phenomena. Indeed, I have now indicated how it can even enable externalists to explain away evidence that might be taken to support internalism as a conceptual claim. It therefore does not seem that the approach taken here can be quickly dismissed merely on the grounds that it offers an externalist account of normative judgment.21

8. Conclusion

In this paper, I have drawn on dual-process models in cognitive psychology to offer a cognitivist, externalist account of the intimate link between normative judgment and motivation. The key innovation in my proposal is that it treats

---

21. My overarching goal in this paper has been to defend cognitivism by showing that externalist versions of the view are more plausible than is commonly thought. However, if my reply to the objection discussed in this section seems unsatisfying, it still might be possible to develop a cognitivist version of weak internalism that is compatible with the dual-process framework sketched above (cf. Tresan 2006; see also Footnote 14).

Ergo • vol. 4, no. 20 • 2017
desire-like, motivationally efficacious states as typical causes of normative beliefs. In Section 5, I argued that this proposal enables cognitivists to explain the alignment pattern as well as noncognitivists can, while doing a better job than noncognitivists in explaining the mismatch pattern. I also argued that even if both views are roughly on par in terms of overall explanatory ability, this is still a serious blow to noncognitivism, since a major part of the appeal of the view was its purportedly unique ability to capture the link between normative judgment and motivation within the Humean framework. And without the support of noncognitivism, expressivism loses much of its appeal as well. This is good news for cognitivists who embrace a truth-conditional approach to normative discourse.

It is important to remember that this good news is only conditional, however. The viability of my proposal will ultimately depend on future work in the science of normative thought and decision making. At the same time, the proposal does enjoy a degree of empirical support, and offers some resources for guiding and interpreting future work.

At the very least, I hope to have shown that the common wisdom regarding the explanatory inadequacy of cognitivism is not a foregone conclusion. The underlying nature of normative thought and discourse might well be just what is suggested by their surface features.

Acknowledgments

For their helpful comments on earlier drafts of this paper, I am grateful to James Beebe, Alex King, Ken Shockley, Lewis Powell, and two anonymous referees for this journal. I am also grateful for helpful discussion with Ariane Nomikos, Neil Sinhababu, Tongli Wu, and audiences at Shandong University and the University at Buffalo.

References


Koenigs, Michael, Liane Young, Ralph Adolphs, Daniel Tranel, Fiery Cushman, Marc Hauser, and Antonio Damasio (2007). Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgments. *Nature*, 446(7138), 908–911. https://doi.org/10.1038/nature05631


Shenhav, Amitai and Joshua D. Greene (2010). Moral Judgments Recruit Domain-General


