JUDGMENT INTERNALISM: AN ARGUMENT FROM SELF-KNOWLEDGE

To be published in *Ethical Theory and Moral Practice* (Volume 21:4, BSET special issue edited by Ben Sachs).

1. INTRODUCTION

This article investigates the relationship between evaluative judgments and motivation. I use the term 'evaluative judgment' to refer to the mental state in virtue of which one counts as sincerely accepting a sentence of the form 'X is F' where 'F' is an evaluative term such as 'good,' 'desirable,' or the like.[1] I thus assume that there are necessary and sufficient conditions for genuinely accepting the sentence 'giving money to charity is good' and that satisfying these conditions requires being in a mental state called 'the evaluative judgment'.

Because I assume a Humean view of human psychology, by 'motivation' I refer to a desire (Smith 1987: 50–54). According to this theory, all mental states are either belief-like cognitive states or desire-like conative states. Belief-like states aim at truth. They have the mind-to-world direction of fit: an essential part of their functional role is to be sensitive to evidence. Furthermore, on their own these states are incapable of moving agents to act. In contrast, the functional role of desire-like states is to make the world fit their content and so desires have the world-to-mind direction of fit. They are dispositional states consisting of being motivated to bring about an outcome and so they can move us to act together with beliefs about the means.

The question then is: what is the relationship between evaluative judgments and desires? If you think that giving money to charity is good, how is that judgment related to your desire to donate money? §2 outlines the internalist and externalist answers to this question and it also explains how this debate's traditional argumentative strategies have led to a standoff.

§3 then suggests that new progress can be made with arguments that have the structure of an inference to the best explanation. As an illustration of this abductive strategy, §3 also outlines a new argument for internalism, which I call 'the Argument from Self-Knowledge'. It suggests that we have reason to believe that internalism is true because its truth would best explain how we are able to use the so-called *transparency method* to know what we desire. The rest of this article defends this new argument. §3–§4 first introduce and

[1] The relevant judgments concern *final, all-things-considered predicative value* rather than attributive, instrumental, prudential, or distinctively moral value (Hurka 2014, §1.4).

motivate the argument's first preliminary premises concerning the type of self-knowledge we have of our desires and the transparency method, which we can use to acquire such knowledge. §5–§7 then focus on the central premise according to which internalism can explain why the transparency method yields self-knowledge better than externalism.

2. Internalism and Externalism

Internalists argue that there is an internal connection between evaluative judgments and desires.[2] Different versions of internalism disagree about the strength of this modal connection. Call the following thesis UNCONDITIONAL INTERNALISM:[3]

> Necessarily, if you judge that some state of affairs S is good, then you have at least some non-derivative desire for S to obtain.

It states that you cannot be in the mental state in virtue of which you count as thinking that S is good unless you have at least some non-derivative desire for S to obtain. Thus, it is impossible for you to think that giving money to charity is good if you do not have at least some direct desire to do so. This view posits a strict internal connection between evaluative judgments and desires.

Call the following thesis CONDITIONAL INTERNALISM:[4]

> Necessarily, if you judge that some state of affairs S is good, then you have at least some non-derivative desire for S to obtain or you are practically irrational (suffering from weakness of will, depression, or other paradigmatic forms of practical irrationality).

According to this thesis, evaluative judgments create requirements of rationality to have desires. If you are in the mental state in virtue of which you count as thinking that giving money to charity is good, then being in this state rationally requires you to have at least some non-derivative desire to donate money. As a consequence, when you are rational (i.e.,

[2] For historical internalists, see Brink (1989: 29, footnote 39). For more recent defenders, see Björklund et al (2011) and Miller (2008: 235–236, footnotes 8–10).
[3] See Stevenson (1937: 16) and Nagel (1970: part 1).
[4] See Korsgaard (1986: 13–15), Smith (1994: 61, 143), Wedgwood (2007, §1.3), and van Roojen (2010: 498–500).

disposed to conform to the requirements of rationality), some direct desire to give money to charity follows from your judgment.

This article defends a version of the Argument from Self-Knowledge the conclusion of which is CONDITIONAL INTERNALISM. If we tried to use a similar argument to argue for UNCONDITIONAL INTERNALISM, that argument would entail that we could know what we desire infallibly. Given that there are general reasons not to accept such views of self-knowledge, I will focus here only on CONDITIONAL INTERNALISM.[5]

Externalists, in contrast, deny that there is an internal connection between evaluative judgments and desires. They defend EXTERNALISM:[6]

> If you genuinely judge that some state of affairs S is good and you are rational, you will desire S to obtain only if there is some further contingent fact about you – *other than your rationality* – that links your judgment to a desire.

On this view, if you judge that giving money to charity is good, it is always a further contingent fact about you whether you also desire to do so. Furthermore, even if you lacked any desire to donate money in this case, you could still count as fully rational.

The debate between internalists and externalists has been intensive because understanding the relationship between evaluative judgments and desires promises to shed light on the nature of evaluative judgments. Consider the Humean arguments for non-cognitivism: the view that evaluative judgments are desire-like attitudes.[7] These arguments have three premises: (i) UNCONDITIONAL INTERNALISM (there is a necessary connection between evaluative judgments and desires), (ii) Humean psychology (any mental state is either belief-like or desire-like), and (iii) Hume's dictum (there are no necessary connections between distinct existences). If (ii) and (iii) are true, then UNCONDITIONAL INTERNALISM entails that evaluative judgments are desires.

Likewise, if you believe that CONDITIONAL INTERNALISM is true, you can consider what evaluative judgments would need to be like in order to create requirements of rationality for having the corresponding desires.[8] Michael Smith argues that for this reason these judgments must be beliefs about which outcomes fully rational versions of ourselves would

[5] See Fernández (2013: 18) and Smith (1987: 46).
[6] Mill and Ross are often thought to have been externalists (Korsgaard 1986: 9). Prominent contemporary externalists include Brink (1989), Shafer-Landau (2003), and Svavarsdottir (1999).
[7] See Blackburn (1984: 188).
[8] Both naturalists (Smith 1994) and non-naturalists (Wedgwood 2007) have pursued this project.

want us to bring about (Smith 1994: ch. 5). Finally, EXTERNALISM naturally leads to cognitivism: the view that evaluative judgments are motivationally inert beliefs.

The debate between internalists and externalists has reached a standoff. Two argumentative strategies have proven inconclusive. The first is called 'the method of cases'. It tries to locate patterns from individual cases by considering hypothetical scenarios involving agents, utterances, actions, and attitudes. We first consult our intuitions about whether agents have made genuine evaluative judgments in these situations and we then reflect whether EXTERNALISM or some form of internalism better makes sense of our reactions.[9]

Internalists focus on cases in which someone changes their mind and utters, for example, that 'giving money to charity is good after all'. If this person has no inclination to donate money when she could easily do so, we think that she is being hypocritical.[10] She didn't genuinely come to think that giving money to charity is good. This intuitive reaction seems to fit internalism better. Externalists, in contrast, focus on amoral, listless and evil people, who seem to make evaluative judgments even if they lack any desire to act accordingly.[11] Externalists then claim that all forms of internalism must be false because they deny that these groups can be both rational and making genuine evaluative judgments.

Using this method is problematic because internalists and externalists have different intuitions about the cases and they also describe them differently. Consider St. Augustine's confession that he once stole pears because doing so was bad (Björnsson 2002: 339). He had no interest in the pears themselves as he could have got better ones in more acceptable ways. According to externalists, St. Augustine made a genuine evaluative judgment (stealing pears is bad), but instead of having the corresponding desire not to steal he desired *to* steal. This is presented as a counter-example to internalism.

Internalists deny that this description of the case is intelligible (ibid.). They claim that St. Augustine had some desire not to steal because we need the motivational pull of his judgment to explain why pursuing the forbidden fruit was so tempting. Described in this way the case supports internalism. Both sides then believe that their descriptions of the case are accurate, and for this reason cases like this cannot be used to make progress.

[9] Sometimes this method is applied by considering the utterances of which individuals and communities we would intuitively translate to our own evaluative language. For internalist formulations, see Hare (1952: 148–149) and Horgan & Timmons (1991). For externalist responses, see, e.g., Dowell (2016).
[10] See Björnsson (2002: 334) and Smith (1994: 6–7, 60). For an externalist response, see Copp (2001: 12).
[11] For lists of discussions, see Miller (2008: 235–236, footnotes 8–10).

The second debated argument is the fetishism argument (Smith 1994: 73–75). It begins from the observation that usually when a person changes her mind about what is good her desires change too. In order to explain this, externalists must claim that the person has a *de dicto* desire with the content 'that I bring about a good outcome'. This desire is needed to explain how you acquire a new desire for giving money to charity when you come to believe that doing so is good. However, Smith points out that this new desire is merely a derivative desire. It would not reflect our genuine direct concern for other people, which is why he thinks that there is something wrong with externalism.[12]

In response, externalists deny that there is anything objectionable about having the *de dicto* desire. Such a standing desire allegedly 'serves as a limiting condition on the formation of other desires', which matches the motivations of a virtuous agent (Shafer-Landau 2003: 159). Externalists also argue that, even if people have the *de dicto* desire and instrumental desires that result from it, they can also care about things directly. As a consequence, externalists see no reason to give up their view because of Smith's argument either.[13]

3. The Argument from Self-Knowledge

I aim to show that we can construct arguments in this debate based on general philosophical premises that do not rely on any controversial intuitions like the previous arguments. One class of intuition-independent arguments have the structure of *an inference to the best explanation*. For the purposes of such abductive reasoning, the internalists should search for some actual facts the obtaining of which their theory can explain better than EXTERNALISM. If there are such facts, they provide an intuition-independent reason for accepting a form of internalism as we should prefer theories on the grounds of their superior explanatory power.

Consider then the following argument:

THE ARGUMENT FROM SELF-KNOWLEDGE

SELF-KNOWLEDGE: Rational agents sometimes know what they desire and some such self-knowledge has two special features: (i) special access (ii) and strong access.

[12] The fetishism argument is a move in the context of the broader best explanation arguments discussed below. The difference between this argument and mine is that, in the fetishism argument, the explanandum is dependent on our moral intuitions about how good people are motivated whereas in my argument it is not based on such controversial intuitions.

[13] See Copp (1997: 48–51), Cuneo (1999), Dreier (2000), Shafer-Landau (2003: 158–159), and Svavarsdottir (1999: §6).

TRANSPARENCY:	Rational agents can acquire self-knowledge of their desires with the features (i) and (ii) by using the transparency method.
EXPLANATION:	CONDITIONAL INTERNALISM explains why rational agents can use the transparency method to acquire self-knowledge of their desires with the features (i) and (ii) better than EXTERNALISM.
CONCLUSION$_1$:	TRANSPARENCY and EXPLANATION together entail that CONDITIONAL INTERNALISM best explains some of the self-knowledge referred to in SELF-KNOWLEDGE.
CONCLUSION$_2$:	CONCLUSION$_1$ gives us a good reason to believe that CONDITIONAL INTERNALISM is true.

SELF-KNOWLEDGE claims that, at least when we are rational, we can know what we presently desire. In some cases we have true beliefs about our own present desires, and these second-order beliefs are sufficiently sensitive to what we actually desire to count as knowledge. They are *safe* in the sense that they could not have been easily false (Byrne 2005: 96–98).[14] SELF-KNOWLEDGE also claims that some of the previous type of self-knowledge of our desires is not based on reasoning that begins from observations of behaviour. Call this the *special access* feature. Furthermore, it claims that often our beliefs about our own desires are more justified than other people's beliefs about them. This leaves room both (i) for us having some false beliefs about our desires and (ii) for other people knowing better in some special circumstances. Call this the *strong access* feature.

There are two strong reasons for accepting SELF-KNOWLEDGE. First of all, this premise has been widely accepted and defended in the self-knowledge literature.[15] Secondly, and more importantly, denying it has significant theoretical costs. If SELF-KNOWLEDGE were false, rational agents could know what they desire only through observing their own behaviour and their own beliefs about their desires would therefore be no more justified than other people's. This would make their lives more difficult (Shoemaker 1996, 26–39). Firstly, when rational agents pursue a shared goal, they must be able to communicate their beliefs and desires to others. For this purpose, they must know in advance what they believe, desire, and intend. Thus, if rational agents did not know their attitudes before they acted, social co-operation could not run smoothly. Furthermore, without advance knowledge of their desires,

[14] I follow Duncan Pritchard's understanding of safety (2009: 34–35): a subject's belief is safe if and only if in the close possible worlds in which the subject continues to form the relevant belief via the same belief-formation method as in the actual world the belief continues to be true.
[15] See Boyle (2009: 136), Byrne (2005: 80–81 and 2011: 202), Fernández (2013: 5–6), and Moran (2001: 10).

rational agents could not form plans that satisfy a combination of their different desires effectively. This means that, if you reject SELF-KNOWLEDGE, then you need to be able to explain how co-operation and planning would be possible without the kind of self-knowledge that SELF-KNOWLEDGE attributes to us.

§4 below explains with examples how the transparency method of TRANSPARENCY functions. These examples will also be used to provide intuitive support for TRANSPARENCY – the claim that the transparency method yields self-knowledge of our own desires. I should also emphasise that this premise too has been widely accepted and defended in the debates about self-knowledge.[16] After the next section on TRANSPARENCY, I will then focus on EXPLANATION, which is the key premise of the Argument from Self-Knowledge.

4. The Transparency Method

TRANSPARENCY is the claim that the transparency method can provide rational agents self-knowledge of their own desires. Let us approach this method from how we can know what we believe. Ask yourself: do you believe that it will rain tonight? The transparency method is the idea that you do not answer this question by turning your attention inwards to the content of your mind but rather you look outwards – you attend to the sky. You answer the question of whether you believe that it will rain by answering the question 'Will it rain tonight?'. More generally, the defenders of the transparency method claim that the way to answer any first-personal question 'Do I believe that P?' is to answer the outward-directed question 'P or not-P?'. After you answer the latter question, you will know what you believe.

Many philosophers think that we can use a similar transparency method to know what we desire.[17] According to them, when we want to know what we desire, we do not turn our attention inwards to the content of our minds but rather we again focus our attention outwards to the qualities of different states of affairs and through this we come to know what we desire.

Imagine that you are having breakfast at a hotel. The buffet offers you different options: you could have fruit, cereal, toast or porridge. Ask yourself: what do you desire to have for breakfast? When we answer this question by using the transparency method, we first consider the qualities of the different options: how healthy they are, whether they fill us up,

[16] The tradition begins from Moore (1903: 446). For the first contemporary defences, see Edgley (1969: 90) and Evans (1982: 224–235). See also Ashwell (2013), Boyle (2009 and 2011), Byrne (2005 and 2011), Burge (1996), Fernández (2013), Gallois (1996), and Moran (2001: esp. 60–62).

[17] See Ashwell (2013), Blackburn (1998: 253–255), Byrne (2005: 99–100 and 2011: 213), Gallois (1996), and Shoemaker (1996: 47).

whether they taste nice and so on.[18] Considering these basic qualities of our alternatives will not, however, be enough to tell us what we desire. Merely by coming, for example, to the conclusion that fruit would be healthy you can only come to know that you believe that the fruit are healthy (Byrne 2005: 99). In order to know whether you have a desire to eat them, you must also consider other questions.

We can see what these additional questions are if we consider just why you would focus on the previous basic qualities of your alternatives. Presumably you consider them because they are the good-making qualities of those alternatives. You would be thinking of these features as their 'merits'. In this way, you are trying to conclude which of your alternatives would be the best choice overall.[19]

This suggests that, more generally, when we use the transparency method for forming beliefs about what we desire we are making evaluative judgments – we are considering how good different outcomes are. We then conclude that we desire an outcome S to obtain on the basis of judging that S is good. If we conclude that S is the best outcome available for us, we know that this is what we want most. If we judge that S is good to a degree even if it not the best alternative overall, we know that we have at least some desire for S to obtain. Therefore, the transparency method for answering the question 'Do I desire S?' relies on considering the question 'Is S good?' and concluding what you desire on that basis. By reaching the conclusion that toast would be the best option, you come to know what you want for breakfast. TRANSPARENCY then says that, by using this method, at least when we are rational we can acquire self-knowledge of our desires with the special and strong access features.

5. Internalist Explanations of the Transparency Method

I have introduced the Argument from Self-Knowledge as a new argument for CONDITIONAL INTERNALISM and explained and motivated its first two premises SELF-KNOWLEDGE and TRANSPARENCY. In the rest, I focus on arguing for the crucial EXPLANATION premise. This section explains how CONDITIONAL INTERNALISM plays an essential role in explaining why the transparency method yields self-knowledge. It focuses on three philosophical explanations of why this method yields self-knowledge of our desires and how CONDITIONAL INTERNALISM naturally supports these explanations.

[18] You might also consider how hungry you are or whether you feel heavy. Crucially, by considering these features of yourself you are not searching introspectively what desires you have.

[19] This type of reasoning is often automatic. You might see the options available to you and recognize one of them as more appealing than others. However, all we need is that at least sometimes rational agents explicitly reason in the described way.

5.1 Rationalism

Rationalists emphasise that we are active thinkers, which gives us an epistemic authority over our own beliefs.[20] We do not just discover that we have certain beliefs, but rather we are responsible for our beliefs because we actively form and revise them on the basis of reasons. We make our minds up by considering evidence in accordance with the central norms of reasoning.

Imagine again that you are asked 'Do you believe that it will rain tonight?'. According to rationalists, because you are responsible for your beliefs in virtue of your rational self-control, you would answer this question by making a decision on the basis of the best reasons available for you. You might, for example, commit yourself to believing that it will rain because of the murky sky. Rationalists then assume that, if you actively form a belief in critical reflection by using the transparency method and you are conceptually competent, you will know what you believe in the same direct way as through raising your arm and conceptual competence you know that you have raised your arm. On this view, questions about our own beliefs are to be understood in a practical spirit rather than in theoretical one, because they call for a resolution rather than a discovery.[21]

Rationalists also believe that, as critical reasoners, we do not just discover our desires, but rather we are responsible for our desires because we have rational control over them in practical deliberation (Moran 2001: 118–119). We form and revise our desires on the basis of the reasons we recognise. In some cases we just have brute urges: we find ourselves desiring certain things. However, usually we make our minds up about what to desire on the basis of reasons. Rationalists claim that, in these contexts, we answer the question 'What do I desire?' with a practical commitment rather than a discovery. They then use the idea that, when we use the transparency method, we actively form our desires by considering reasons to explain why the resulting beliefs about our desires count as knowledge. We know that we have certain desires because we have formed them ourselves whilst using the transparency method.

How does this explanation relate to CONDITIONAL INTERNALISM? When we as rational agents use the transparency method, we come to know what we desire by considering what is good. If we really are *forming* desires in this process like the rationalists insist, then thinking about what is good must be for rational agents a way of actively forming and revising desires

[20] See Boyle (2009 and 2011), Burge (1996), and Moran (2001: §2.5–§2.6).

[21] Just how this outline of an explanation (following Fernández (2013: 20)) translates to a full theory about how self-knowledge is achieved is controversial. See Burge (1996: 100–115), Fernández (2013: 18–19), Gertler (2011: §6.3–§6.4), and Moran (2003: 405–406).

(Moran 2001, 57). It is then essential to notice that thinking about what is good is an excellent way of forming desires if the evaluative judgments in question are either desires themselves or they have a direct, unmediated power to produce desires in us insofar as we are rational. In either case, there would be an internal necessary connection between evaluative judgments and the desires of rational agents just as the internalists claim. The rationalist explanation for why the transparency method can yield self-knowledge about desires therefore works naturally if CONDITIONAL INTERNALISM is true. CONDITIONAL INTERNALISM offers the rationalists the missing explanation of why thinking about what is good is an excellent way of forming desires.

5.2 Byrne's Rule of Reasoning

Alex Byrne offers an alternative explanation of why using the transparency method yields self-knowledge (Byrne 2005 and 2011). He claims that we should understand the transparency method as an epistemic rule we can explicitly follow in our reasoning.[22] In the case of beliefs, it tells us to 'If P, believe that you believe P!' (Byrne 2005: 95). Byrne then observes that you explicitly follow this rule only if you do what the consequent tells you to do *when and because you think that the antecedent is true*. His main thesis is that using the rule in this way is *self-verifying* (Byrne 2005: 96; Byrne 2011: 206–207). Whenever you think that the antecedent is true you thereby believe that P and therefore your belief that you believe that P will generally be true when you use the rule.[23] For example, you see dark clouds, which suggest to you that it will rain tonight. If you explicitly follow Byrne's rule by starting from thinking that the antecedent is true (which consists of thinking that it will rain tonight), the resulting belief that you believe that it will rain tonight will be true.

A similar explanation can be offered in the context of desires. Consider the following transparency rule of reasoning: "If state of affairs S is good, believe that you desire S to obtain!" You explicitly follow this rule only if you do what the consequent tells you to do (believe that you desire S to obtain) *when and because* you think that the antecedent is true (you think that S is good).

Reasoning in accord with this rule is self-verifying for rational agents *if CONDITIONAL INTERNALISM is true*. Thinking that the antecedent of the rule is true requires making an evaluative judgment, thinking that S is good. If CONDITIONAL INTERNALISM is true, there is an internal necessary connection between such evaluative judgments and desires in rational

[22] For epistemic rules, see Byrne (2005: §7.1).
[23] For a critical examination, see Ashwell (2013) and Boyle (2011).

agents. An evaluative judgment that S is good makes it the case that you desire S to obtain either because the evaluative judgment itself is a desire or because it has a direct, unmediated power to produce desires in you insofar as you are a rational agent. This feature of CONDITIONAL INTERNALISM explains why, when rational agents do what the consequent tells them to do (believe that they desire S to obtain) on the basis of recognising that the antecedent is true (an evaluative judgment that S is good), they acquire a true belief about their own desire. Therefore, if CONDITIONAL INTERNALISM is true, following the rule itself makes it the case that rational agents will have self-knowledge of your their desires, which is why using this rule is self-verifying for them. This is why the truth of CONDITIONAL INTERNALISM supports Byrne's explanation of why the transparency method yields self-knowledge of our desires.

5.3 The Bypass View

The third explanation of why the transparency method yields self-knowledge is the bypass view (Fernández 2013: ch. 2). We form our beliefs on the basis of experiences, memories, testimony and other beliefs. For example, you might form the belief that there is an apple in front of you on the basis of seeing one. In order for this visual experience to support your belief about the apple, there must be a correlation between (i) having the visual experience of seeing an apple and (ii) there being an apple in front of you.

According to Fernández (2013: 49), when you use the transparency method, you form the second-order belief that you have a certain first-order belief on the basis of the same state that is your ground for the relevant first-order belief. For example, the transparency method suggests that it is natural to use the perceptual experience of seeing an apple (P) both as a ground for believing that there is an apple in front of you (belief that P) and for the second-order belief that you have that first-order belief (belief that you believe that P).

In order for us to be justified to form our second-order beliefs in this way, a reliable correlation is required between being in the given grounding state and having the relevant first-order belief. Such a correlation is enough to guarantee that being in the grounding state also correlates with the truth of the relevant second-order belief. Hence, as long as we form beliefs uniformly on good grounds, the beliefs about our own beliefs we form by using the transparency method will count as self-knowledge.

Fernández (2013: ch. 3) applies his bypass model also to desires. He begins from the idea that we form our desires on the basis of other experiences, urges, desires, beliefs and judgments. For example, you might form a desire to eat on the basis of feeling hungry. Let

us call these states on the basis of which we form desires the 'grounds' of those desires (Fernández 2013: 82; Nagel 1970: 29). Evaluative judgments too are grounds of desires in this sense. We often form a desire for the state of affairs S to obtain on the basis of thinking that S is good. When we use the transparency method, we form beliefs about our own desires on the basis of the grounds for those desires (Fernández 2013: 86). For example, it is natural to use the judgment that giving money to charity is good both as a ground for desiring to give money to charity and as a ground for the higher-order belief that you have that desire.

What then entitles you to assume that the support you have for believing that you desire S can be the same evaluative judgment that supports your desire for S to obtain? All that is required to justify this assumption is that there is a reliable correlation between being in the given grounding state (evaluative judgment) and having the desire in question (Fernández 2013: 87–91). This correlation is enough to guarantee that being in the relevant grounding state also correlates with the truth of the relevant higher-order belief. Thus, as long as we form desires on the basis of evaluative judgments, forming our higher-order beliefs on those same grounds gives us knowledge of our own desires.

The truth of CONDITIONAL INTERNALISM would explain why there would be this reliable correlation between our evaluative judgments and our desires when we are rational. Internalists, after all, argue that there is an internal necessary connection between evaluative judgments and desires either because evaluative judgments themselves are desires or because they have a direct, immediate power to produce desires in us insofar as we are rational agents. This is why the internalists can endorse Fernández's bypass explanation of why the transparency method yields self-knowledge of our desires. They can provide Fernández the required modal link between the evaluative judgments and desires of rational agents that justifies the higher-order beliefs about their own desires, which rational agents form on the basis of the evaluative judgments when they use the transparency method. This is why INTERNALISM supports also Fernández's explanation of why the transparency method yields self-knowledge of our desires.

This section has shown how CONDITIONAL INTERNALISM naturally supports the three best philosophical explanations of why the transparency method yields self-knowledge of our desires. Recall that EXPLANATION claims that CONDITIONAL INTERNALISM can explain why the transparency method yields self-knowledge of our desires *better than* EXTERNALISM. This section has then argued only for one half of this premise: that CONDITIONAL INTERNALISM explains well why the transparency method yields self-knowledge of our desires when we are

rational. The next section will argue for the second half: the externalists cannot explain why the transparency method works equally well.

6. Externalist Explanations of the Transparency Method

Externalists have at least the following three potential ways to explain how the transparency method yields self-knowledge of our desires:

(i) Externalists could explain why making evaluative judgments would be a good way of forming desires even if EXTERNALISM were true and so vindicate the rationalist account of the transparency method.

(ii) Externalists could explain why using Byrne's transparency rule of reasoning would self-verifying even if EXTERNALISM were true and so vindicate Byrne's self-verification account of the transparency method.

(iii) Externalists could explain how there could be a reliable correlation between evaluative judgments and desires even if EXTERNALISM were true and so vindicate the bypass account of the transparency method.

All these strategies require the same: showing that there could be a strong enough correlation between evaluative judgments and desires even if EXTERNALISM were true. Such a correlation would make evaluative judgments (a) a good way of forming desires, (b) Byrne's transparency rule self-verifying, and (c) justify using evaluative judgments as grounds for both desires and higher-order beliefs about those desires. Thus, all the externalists need is an account of a strong correlation between evaluative judgments and desires that is compatible with EXTERNALISM. I will first consider correlations grounded on biological and cultural facts and then ones grounded on *de dicto* desires for bringing about good outcomes.

6.1 Biological and Cultural Facts

In order to create a sufficiently strong correlation between our evaluative judgments and desires, externalists could first argue that, due to (ii) contingent biological facts about human psychology or (ii) contingent cultural facts, people who make evaluative judgments tend to have the corresponding desires (Shafer-Landau 2003: 159). Externalists could then claim that these biological and cultural links are enough to explain why using the transparency method yields self-knowledge of our desires.

The problem is that externalists have only suggested that there might be some biological or cultural facts that create a correlation between evaluative judgments and desires, but they have not described what those facts are. Because of this, we cannot judge what kind

of correlations between our evaluative judgments and desires these additional facts might produce and whether these correlations would be strong enough to support the philosophical explanations of why the transparency method yields self-knowledge of our desires when we are rational. This means that currently the internalist explanation of why the transparency method yields self-knowledge is the best explanation because it is the *only* explanation.

Therefore, at least the outlined argument presents the externalists with a challenge: they have to be able to produce as good explanations (based on some specific actual biological or cultural facts) of why the transparency method yields self-knowledge as the internalist explanations outline above. Until externalists have provided these explanations, the Argument from Self-Knowledge gives us a reason to prefer CONDITIONAL INTERNALISM. Currently then, CONDITIONAL INTERNALISM has an advantage because it can explain something that EXTERNALISM leaves mysterious: how we can know what we desire by thinking about what is good.

This challenge will not be trivial. The externalist accounts will have to explain why the self-knowledge rational agents gain by using the transparency method is *safe*: why the beliefs about their own desires formed by thinking about what is good could not have been easily false. Here we need to consider the closest worlds in which the biological and cultural facts are just slightly different from the ones used in the real externalist explanations. The concern is that in those worlds rational agents could be using the transparency method to form false beliefs about their desires without noticing, which will threaten the safety of the true beliefs formed with the transparency method in the actual world. The externalists will insist that the previous worlds are sufficiently far away from the actual world so as not to threaten the safety of the actual beliefs. However, for this response to work, the externalists would need to describe what the relevant biological and cultural facts are as only then they would be able to give an account of just why these worlds are not close enough to the actual world to be relevant for safety. Without this type of an account, the externalist explanations of the transparency method threaten to leave the safety of our self-knowledge of our desires unexplained.

6.2 De Dicto Desires

The externalists could also return to their explanation of how our desires change when we make evaluative judgments. This explanation is based on the *de dicto* desire with the content 'that I bring about a good outcome' (§1). According to externalists, this desire – together with a belief that an outcome is good – leads to a derivative desire for bringing about the relevant

outcome. Externalists could then argue that the existence of the background *de dicto* desire for good outcomes is sufficient for creating a reliable correlation between our evaluative judgments and the corresponding desires. This correlation would be enough to explain why the transparency method yields self-knowledge of our desires.

This proposal too faces a challenge.[24] Let us stipulate that Daniel is an actual agent, who has the relevant *de dicto* desire and who forms true beliefs about his own desires with the transparency method. There is then a range of close possible worlds in which Daniel lacks the *de dicto* desire and in which he is also inclined to use the transparency method. In order for Daniel's actual beliefs about his desires to count as self-knowledge, these beliefs must be safe. This entails that, in the close worlds in which Daniel lacks the *de dicto* desire, something must alert him of the fact that, if he were to use the transparency method, he would form false beliefs. The only potential warning-sign then seems to be awareness of the existence of the *de dicto* desire itself. If Daniel were able to easily know whether he has this desire in the relevant close worlds, he could use this self-knowledge to determine whether he can use the transparency method to acquire self-knowledge of his other desires in those worlds. The question then is: how could Daniel know whether he has the relevant *de dicto* desire?

The externalists cannot claim that Daniel could come know that he has the relevant *de dicto* desire by using the transparency method itself. After all, Daniel is considering in a range of close possible worlds whether he is able to use the transparency method to form true beliefs about his desires. The externalists also cannot claim that Daniel could know that he has this desire on the basis of observing his own behaviour. In that case, Daniel would be unable to know whether he is entitled to use the transparency method before he has observed his own behaviour. This means that Daniel would be unable to acquire self-knowledge with the special access feature (§2). The externalists then owe us an alternative explanation of how we can know whether we have the relevant *de dicto* desire for bringing about good outcomes under that description.

7. Alternative Accounts of Self-Knowledge

In addition to the transparency method, there are two other orthodox views of self-knowledge.[25] Firstly, externalists could attempt to defend acquaintance theories of self-

[24] These same problems also apply to the externalist higher-order desire strategy (see Dreier 2000).
[25] Here I follow Gertler (2011).

knowledge.[26] When applied to desires, they claim that desires are 'luminous' or 'self-presenting' – they have a certain phenomenal feel as a part of their essence.[27] On this view, there is something it is like to have desires, and thus a direct awareness of a desire is built into the desire itself. We are then supposed to know what we desire because we stand in the metaphysical relation of immediate acquaintance to our desires.

Some desires admittedly have a phenomenal feel to them (Russell 1912: 77; Smith 1987: 45). There is something it is like to desire food when you are hungry. However, not all desires have a special phenomenology. Many calm desires are dispositional states that play a role in the causal explanations of actions even if we are not immediately aware of them (Smith 1987: 46–47). Consider Smith's example: John is pursuing a career in arts, but ceases to do so after his mother's death. Previously John thought that he is pursuing this career for the sake of art itself, but now he comes to recognize that he was motivated by a desire to please his mother. It seems plausible to ascribe this desire to John in order to explain his previous behaviour even if he never had immediate awareness of it. This means that at best the acquaintance theory can explain how we know about a limited class of desires. It also means that externalists still need to explain in some other way how we can acquire self-knowledge of our calm dispositional desires such as the *de dicto* desire with the content 'that I bring about a good outcome'.

Externalists could also attempt to defend inner sense theories of self-knowledge, which attempt to understand self-knowledge with the model of perception.[28] Perhaps in addition to our outer senses, we also have a special faculty of an inner sense. It can be understood as the brain's self-scanning mechanism, which is causally sensitive to the target beliefs and desires in the same way as our ears are sensitive to sounds. The inputs to this mechanism are our beliefs and desires and the outputs higher-order beliefs about those inputs. The reliability of our self-knowledge is then explained by the idea that the output states of introspection are causally sensitive to the input beliefs and desires they are about.

There are many well-known objections to this theory of self-knowledge.[29] Firstly, the inner-sense theories describe a sub-personal mechanism, a brain-level scanning mechanism that is causally responsible for the production of our second-order beliefs. However, on the personal level of deliberation, when that mechanism causes us to have second-order beliefs

[26] For contemporary defenders, see Gertler (2011: 94).
[27] See Platts (1979: 256).
[28] For contemporary defenders, see Gertler (2011: 132).
[29] For an overview, see Gertler (2011: §5.3–§5.5). For additional objections, see Boghossian (1989), Burge (1996: 105) and Moran (2001: 13–15).

there would be no evidence or reasons we can cite in support of those beliefs (Fernández 2013: 28–36; Peacocke 1999: 224–225). This is because the functioning of the mechanism itself is inaccessible to us in reflection. As a consequence, the inner-sense theories threaten to leave us in the position of 'the infallible psychic who just finds herself believing things about the future for no good reason' (Zimmerman 2006, 349). The critics argue that, when it comes to self-knowledge, we simply are not in this position: there are things we can say to support our self-knowledge.[30]

The second objection begins from the intuitive idea that self-knowledge is necessarily asymmetrical (Gertler 2002: §2). The way we know what we believe and desire is necessarily such that others could not use that same method to know what we believe and desire. It is then argued that the inner-sense theories can only support contingent asymmetry. As it happens, our self-scanning mechanisms are wired so that our own beliefs and desires reliably cause higher-order beliefs only in us. However, if the inner-sense views were true, nothing would rule out creating a similar scanning mechanism that would connect causally your beliefs and desires to my brain and thus give me the same knowledge of your mental states as you have. The critics of the view argue that the inner sense views must be defective because they leave room for this possibility.

8. Conclusion

The aim of this article was to show that we can make new progress in one of the most important metaethical debates that has reached a standoff. I suggested that we cannot make such progress as long as we rely on strategies that are based on controversial intuitions about individual cases or virtuous agents. I then argued that the way forward is to consider what actual facts the truth of different forms of internalism could explain best. The abductive reasoning to the best explanation of those facts would provide an intuition-independent argument for the internalist theories.

As an illustration of this strategy, I outlined the Argument from Self-Knowledge with three premises: rational agents sometimes know what they desire and some of this knowledge has the special and strong access features (SELF-KNOWLEDGE); rational agents can get such self-knowledge by using the transparency method (TRANSPARENCY); and CONDITIONAL INTERNALISM explains why the transparency method yields self-knowledge of desires better than EXTERNALISM (EXPLANATION). The rest of this article then defended this argument. §3–

[30] These critics claim that, for this reason, the inner-sense views must be committed themselves to implausible forms of externalism in epistemology (Peacocke 1999: 241).

§4 explained and motivated the first two general preliminary premises SELF-KNOWLEDGE and TRANSPARENCY. After this, the article focused on arguing for the key premise EXPLANATION. This means that, if my arguments in §5–§7 for EXPLANATION work, an intuition-independent argument for CONDITIONAL INTERNALISM can be given by defending SELF-KNOWLEDGE and TRANSPARENCY in the philosophy of self-knowledge where these claims have already been widely accepted.

References

Ashwell, L. (2013). Deep, Dark … or Transparent: Knowing Our Desires. *Philosophical Studies*, 165(1), 245–256.

Björklund, F. et al. (2011). Recent Work: Motivational Internalism. *Analysis*, 72(1), 124–137.

Björnsson, G. (2002). How Emotivism Survives Immoralists, Irrationality, and Depression. *The Southern Journal of Philosophy*, 40(3), 327–344.

Blackburn, S. (1984). *Spreading the Word*. Oxford: Oxford University Press.

Blackburn, S. (1998). *Ruling Passions*. Oxford: Oxford University Press.

Boghossian, P. (1989). Content and Self-Knowledge. *Philosophical Topics*, 17(1), 65–82.

Boyle, M. (2009). Two Kinds of Self-Knowledge. *Philosophy and Phenomenological Research*, 78(1), 133–163.

Boyle, M. (2011). Transparent Self-Knowledge. *Proceedings of the Aristotelian Society*, Supplement, 85, 223–241.

Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Burge, T. (1996). Our Entitlement to Self-Knowledge. *Proceedings of the Aristotelian Society*, 96, 91–116.

Byrne, A. (2005). Introspection. *Philosophical Topics*, 33(1), 79–104.

Byrne, A. (2011). Self-Knowledge and Transparency. *Proceedings of the Aristotelian Society*, Supplement, 85, 201–221.

Copp, D. (1997). Belief, Reason and Motivation: Michael Smith's *The Moral Problem*. *Ethics*, 108(1), 33–54.

Copp, D. (2001). Realist Expressivism: A Neglected Option for Moral Realism. *Social Philosophy and Policy*, 18(2), 1–43.

Cuneo, T. (1999). An Externalist Solution to the 'Moral Problem'. *Philosophy and Phenomenological Research*, 59(2), 359–380.

Dowell, J. (2016): The Metaethical Insignificance of Moral Twin Earth. *Oxford Studies in Metaethics*, 11: 1–27.

Dreier, J. (2000). Dispositions and Fetishes: Externalist Models of Moral Motivation. *Philosophy and Phenomenal Research*, 61(3), 619–638.

Edgley, R. (1969). *Reason in Theory and Practice*. London: Hutchinson.

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

Fernández, J. (2013). *Transparent Minds*. Oxford: Oxford University Press.

Gallois, Á. (1996). *The World Without, The Mind Within*. Cambridge: Cambridge University Press.

Gertler, B. (2002). The Mechanics of Self-Knowledge. *Philosophical Topics*, 28(2), 125–146.

Gertler, B. (2011). *Self-Knowledge*. London: Routledge.

Hare, R.M. (1952): *The Language of Morals*. Oxford: Oxford University Press.

Horgan, T. & Timmons, M. (1991): New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*, 16: 447–465.

Hurka, T. (2014). *British Ethical Theorists from Sidgwick to Ewing*. Oxford: Oxford University Press.

Korsgaard, C. (1986). Scepticism about Practical Reason. *Journal of Philosophy*, 83(1), 5–25.

Miller, C. (2008). Motivational Internalism. *Philosophical Studies*, 139(2), 233–255.

Moore, G.E. (1903). The Refutation of Idealism. *Mind*, 12(48), 433–453.

Moran, R. (2001). *Authority and Estrangement*. Princeton: Princeton University Press.

Moran, R. (2003). Responses to O'Brien and Shoemaker. *European Journal of Philosophy*, 11(3), 402–419.

Nagel, T. (1970). *The Possibility of Altruism*. Princeton: Princeton University Press.

Peacocke, C. (1999). *Being Known*. Oxford: Oxford University Press.

Platts, M. (1979). *Ways of Meaning*. London: Routledge.

Pritchard, D. (2009). *Knowledge*. Houndmills: Palgrave MacMillan.

Russell, B. (1912). *The Problems of Philosophy*. New York: Henry Holt & Co.

Shafer-Landau, R. (2003). *Moral Realism – a Defence*. Oxford: Oxford University Press.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Smith, M. (1987). Humean Theory of Motivation. *Mind* 96(381), 36–61.

Smith, M. (1994). *The Moral Problem*. Oxford: Blackwell.

Stevenson, C. (1937). The Emotive Meaning of Ethical Terms. *Mind* 46(181), 14–31.

Svavarsdottir, S. (1999). Moral Cognitivism and Motivation. *The Philosophical Review*, 108(2), 161–219.

Van Roojen, M. (2010). Moral Rationalism and Rational Amoralism. *Ethics*, 120(3), 495–525.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

Zimmermann, A. (2006). Basic Self-Knowledge: Answering Peacocke's Criticism of Constitutivism. *Philosophical Studies*, 128(2), 337–379.