MASTER IN ANALYTIC PHILOSOPHY

TFM/Final Research Paper

September 2024

Transparent Desires and the Uniformity of Self-Knowledge

by Juan Pablo Cortés Bau

Under the supervision of:

Fernando Broncano-Berrocal

Transparent Desires and the Uniformity of Self-Knowledge

Abstract: The transparency method claims that we can gain knowledge of our own minds by considering the world. In particular, it says that a subject can know whether she believes that p by a world-directed question of the form "Is *p* true?". Something similar could be tried regarding desires: I can know what I desire by considering the qualities of the intentional object of desire. Defending this last claim involves formulating a world-directed question for the method and defending that this method possesses a strong degree of epistemic warrant. The most prominent theories of the transparency method applied to desire are the bypass view (Fernández 2008), the desirability rule (Byrne 2018), and the conceptual approach (Andreotta 2020). This paper argues that none of these proposals apply successfully the transparency method to desires. Finally, I argue that the transparency method can be partially applied to desires if we take into consideration the distinction between passive and active self-knowledge (Boyle 2009).

1. Introduction

In the last few decades, the transparency method has gained traction among philosophers concerned with self-knowledge, such as Moran (2001), Boyle (2024), Byrne (2018), Silins (2013), and Valaris (2014). Evans (1982) first proposed this method, and it was first sketched as a method centered on knowing our own beliefs. Evan's basic proposal is that to know whether I believe that *p*, I do not have to look "inside" my mind. Instead, I must look outside, to the world, and consider whether *p* is the case or not. If I can answer the question, "Is *p* true?", then I know whether "I believe that *p*" is true or not. Evans puts it in the following way:

> If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' I get myself in

> a position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p*. (Evans, 2002, p. 225)

According to this approach, I am able to answer a "mental question" about myself by answering a "non-mental" question about the world. Continuing Evans' example, by evaluating the political and economic state of the world, I can have knowledge about my own beliefs regarding a possible war. What must be highlighted here is that the transparency method is characterized by a world-directed question. In this sense, the mental, internal question "Do I believe *p*?" is transparent to the world-directed question "Is *p* the case?".

Evans' proposal is tempting in at least two different respects. As one can see at the end of Evans' quote, the theory only postulates the basic procedures that we use to form first-order beliefs (2002, p. 225). In other words, we do not need to postulate the existence of a mysterious inner sense that would allow us to scan our own minds to gain access to our own mental states[1]. Our means to generate first-order beliefs would be enough to allow us to generate second-order beliefs. For this reason, the transparency method is metaphysically economical (Ashwell, 2013, p. 247).

On the other hand, the method seems very promising when one expects to account for first-person privileged access. The method is able to explain how it is that we have a special epistemic justified path to get knowledge of our own mental states that does not rely on a third-person approach, such as behavioral evidence. In other words, the transparency method assumes the thesis of epistemic asymmetry, which states that a subject "may have a type of warrant for the belief that she is in mental state M which is unavailable to others" (Gertler, 2003, p. 22).

To illustrate this with Evans' example, if I determine that there will be a third world war, I am then justified to say that I believe that there will be a third world war. Notice that this does not mean that my belief "There will be a third world war" is necessarily justified. I could be mistaken, and I could believe that without sufficient reasons to make it a justified belief. But that is not necessary for the method. What is epistemically justified is the assertion

[1] Transparency theorists, including Evans, have attacked repeatedly the defenders of the "inner scan" view. Some of the defenders of this latter view are Armstrong (1968) and Nichols and Stitch (2003).

"I believe that there will be a third world war", even if I believe this because I had a dream in which it happened.

However, I am not justified in saying that Gabriel, my cousin, believes that there will be a third world war using this exact same process. The method is a first-person method, which implies that it is only reliable when someone applies it to oneself. In this sense, the transparency method is always a first-person method of self-knowledge. The transparency method does not allow us to gain knowledge of other people's minds. Notice, however, that I could apply the transparency method to determine whether I believe that "Gabriel believes that there will be a third world war" or not. Nevertheless, this is not a problem since I can judge the truth of this proposition by looking at the world. In this case, I would make a judgment by observing Gabriel's behavior. Gabriel's beliefs are not transparent to me, but my beliefs about Gabriel's beliefs can be transparent, and this is completely coherent with the theory.

There exists an ongoing debate regarding the extent of the transparency method. Many defenders of the transparency method have claimed that it is possible and desirable to extend Evans' theory beyond beliefs, and they have tried to apply it to different mental states (such as Andreotta, 2020; Barz, 2015; and Paul, 2012). Specifically, there have been several attempts to apply the transparency method as a way of getting to know our own desires. The aim of this paper is to examine the three most prominent attempts of applying the transparency method to desires: the bypass view, DES, and the conceptual approach. I will consider each one separately, explaining the main claims of each one, and then I will present some problems that arise from these proposals. I will conclude that none of these proposals are able to give a full, robust theory of the self-knowledge of desires.

Before we start with the actual proposal, it is important to note that this discussion is framed into a bigger discussion regarding the nature of self-knowledge. Boyle (2009, p. 141) has pointed out that underlying the discussion of self-knowledge lays the *Uniformity Assumption.* This assumption, according to Boyle, states that there is just one fundamentally uniform way of explaining every type of self-knowledge that possesses first-person authority[2]. This assumption has become important for the debates regarding the transparency

[2] First person authority (also called the thesis of incorrigibility) is the claim that if someone self-ascribes a mental state, other people cannot reasonably deny or dispute that self-ascription (Gertler, 2003, p. 22). First

method. Some defenders of this method have tried to solve the *Generality Problem*, which is the possibility of extending the transparency method to all, or most, of our propositional attitudes (Andreotta, 2020). Some defenders of the transparency method, such as Andreotta (2020), Boyle (2024), and Byrne (2018), have tried to claim that it is at least plausible to generalize the transparency method and to give a robust account of self-knowledge. Solving the Generality Problem, or developing a theory of transparency that can be generally applied to every propositional attitude, would show that the Uniformity Assumption is true. In fact, most of the researchers that face the Generality Problem presuppose the Uniformity Assumption.

The Uniformity Assumption, when stated explicitly, would become a Uniformity Thesis. This thesis states that we can account for the self-knowledge of all the different propositional attitudes with, mostly, one method. Furthermore, this method can also explain the kind of privileged access present in all the propositional attitudes. In this paper I will also argue that, in light of the theories of transparent desires that have been developed so far, we cannot accept the Uniformity Thesis. The transparency method seems to be a good account for explaining self-knowledge of beliefs, but it seems to lack its explanatory virtues when we turn our attention to desires. If this were the case, we would need an account to explain the privileged access present in beliefs, but a different account to explain the privileged access of desires. Then, there would be at least two accounts of self-knowledge that would complement each other.

Another important point that has to be considered is that among the many debates that form part of the epistemology of self-knowledge, there is the debate about the infallibility thesis. This thesis, in its most basic form, states that if a subject believes that she is in a determinate mental state, then she is in fact in that determinate mental state (Gertler, 2003). The infallibility thesis, at least in this general form, has been widely rejected. There seems to be too many cases in which a subject believes that she is in some mental state and she is mistaken. The most classic example is the psychoanalysis patient. I go to therapy, and my therapist tells me that I am jealous of my brother. I am, in fact, not jealous of him, but my

person authority is one of the theses that take the form of the 'privilege access' that is presupposed as a fundamental part of self-knowledge.

therapist, with her position of epistemic superiority and power, makes me believe that I am jealous. This case would show that the infallibility thesis is false.

Neither Evans nor any other defender of the transparency method, as far as I know, accepts the infallibility thesis. The transparency method leaves room for error. Namely, cases in which I follow the method and I form a false second order belief about myself. However, the consequence of denying the infallibility thesis is that transparency theorists have to specify the type of epistemic justification that the transparency method grants. That is why they speak of *strength* (Fernandez, 2008), for example, as a principle of epistemic privilege, about *safety* (Byrne, 2018) or *reliability* (Roche, 2023) to account for the production of knowledge of the method. The task, then, is very complicated, and it requires very precise theoretical machinery: the method must leave room for error, but it also has to provide a path that leads to knowledge.

If we want to judge the success of a theory of transparent desires, then we need to postulate some basic criteria that the theory has to take into account. I propose three criteria for theoretical adequacy: 1) It has to explain the type of epistemic warrant that we can expect from the method[3], 2) it has to propose a world-directed question as the method of gaining self-knowledge of desires, and 3) it has to apply to all types of desires.

The first criterion has its grounds in the following consideration: when applied to beliefs, the transparency method is not just a rule of thumb that we apply in our daily life when we are not sure about what we believe or what we should believe. The transparency method is so interesting for philosophical debate precisely because it brings to the light a deep and fundamental relation between the world and our beliefs. Even if we still lack a consensus about the general understanding of the transparency method applied to beliefs, we could say that a possible explanation of this relation is that our beliefs tend to track the truth, which allow us to form a strong connection between a world-directed question and our own mental states. Postulating a strong, basic connection between the world and our beliefs allows the transparent theorist to defend that the method actually provides knowledge of our own minds. If someone applying the transparency method comes to believe that *p*, then she has a

[3] Different authors propose different epistemic degrees of warrant. Byrnes (2018), for example, states that his method produces safe beliefs. Andreotta (2020) states that his method is grounded on conceptual entailment, which seems to give a stronger degree of warrant than Byrne's security. Fernández (2018) talks about the strength of the justification of his method. I use the term "warrant" because it is the most neutral term.

very strong justification to believe that she believes that *p*. This epistemic warrant that is present in the transparency of our beliefs has to be maintained if we want to apply it to desires. If it does not, then we would not be able to account for any type of privileged access and the theory would lose one of its most important virtues.

The second criterion is a matter of coherence and continuity. If we can gain knowledge of our own beliefs through a world-directed question, then, if we want to apply the same method to desires, we need to use also a world-directed question. If we use a mental-directed question, for example, we would not be applying the method of transparency; we would be doing something else.

The third criterion comes as an implication of the Uniformity Thesis. If we expect to explain the self-knowledge of every type of propositional attitude with just one basic method, then this method has to be able to account in general for all the different types of desires that there might be.[4] If the transparency method can only account for a specific type of desire, we would need to abandon the Uniformity Thesis. Furthermore, if transparency were not good enough to explain every type of desire, we would most probably need to postulate again some type of "inner sense" or scanner for the unaccounted desires, which would deprive us from the metaphysical economy that transparency seemed to promise.

2. The bypass view

Fernández (2003, 2008) argues that the transparency method can be successfully applied to desires. To defend his claim, he proposes the Bypass View. He first starts with a taxonomy of desire. According to him, there are three fundamental forms of desire. The first one is the distinction between instrumental and non-instrumental desires. Fernández gives the following definition of instrumental desires: "For any propositions p, q and subject S: in normal circumstances, if S desires that p and S believes that p would be the case if q were the case, then S desires that q". (2008, p. 521) Instrumental desires are special, since they depend on another desire. In other words, I only desire *q* because I desire *p*, and *q* would

[4] More about the different types of desire in what comes next.

allow me to get *p*. The grounding of the desire is another desire. It is important to keep this in mind.

Fernández's view is based on the notion of "grounding". According to him, some propositional attitudes tend to cause other propositional attitudes. If propositional attitude B is usually caused by propositional attitude A, then we can say that B is grounded on A. It is not a strong relation of implication, but there is some causal relation between them. This notion of grounding, as we will see later on, is essential for Fernández's account.

Following with the taxonomy, there are two types of non-instrumental desires, the first one being desires that come prompted by urges. According to Fernández, there are certain mental states that are not desires but that are quite related to them. He calls them urges, and they are any type of "appetites, craving, yearnings and longings" (*ibid.* p. 522) considered from a quite general stance. Whenever I feel like doing something, for Fernández, I have an urge. If I feel like having a walk, if I suddenly have the wish to talk to a friend, or if I really want to eat an ice cream: these are all instances of urges. Urges tend to cause "basic" desires. When I urge to eat a burger, then the desire of eating a burger arises in me. The relation between urges and desires is not causal, since, according to Fernández, it is possible to have an urge and not have the corresponding desire. He exemplifies this with the case of a child that is sexually aroused but does not possess the conceptual tools that are necessary to understand what intercourse is. This would be a case, supposedly, of an urge without desire.

The last type of non-instrumental desire is based on value. Fernández states that when someone values something, then that value grounds the desire for that something. For example, if I value studying philosophy, then I desire to study philosophy, and it is the value that I put into the study of philosophy what grounds my desire to do so. These desires are not as "basic" as the desires that come from urges, since "value" is a much more refined terms that allows for rational consideration, and it can contemplate reasons.

From this taxonomy of desire Fernández then provides the "bypass view", which would be a transparent method to achieve self-knowledge. This method aims to explain privileged access, epistemic asymmetry[5], and the fact that the beliefs that result from this

[5] Epistemic asymmetry is the following claim: A subject has a warrant for the belief that she is in a determinate mental state that is not available to other people (Gertler, 2003, p. 21).

method are "strongly justified" (2003, p. 363), in the sense that they would not be susceptible to the type of errors that we might encounter when ascribing mental states to others. The general formulation of the bypass view is the following one:

The bypass view

For any proposition p and subject S:

Normally, if S believes that she wants that p, then there is a state E such that

(a) S's belief has been formed on the basis of her being in E.

(b) E constitutes grounds for the desire that p in S. (Fernández, 2008, p. 524)

With this formulation, according to Fernández, we would be able to account for all the different types of desires that we might possess. It is called the "bypass" view because we do not consider the desire in question to know whether we possess such a desire, but we look beyond the desire and look at what grounds it. This intention of looking beyond would be, according to Fernández, what makes this method a transparency method. If I want to take the bus to get to the airport, to follow Fernández's example, I could know that I desire this because I am able to know that I possess the desire of going to the airport, and taking the bus would allow me to go to the airport. I can know my desire of getting the bus because it is grounded on my desire of getting to the airport and my belief that the bus will take me to the airport. This would be a case of instrumental desire, and instrumental desires are grounded in a desire-belief pair.

Fernández's example illustrates his interpretation of (a). Forming a belief *on the basis* of being in E means that the subject is aware of that state before forming the relevant belief. I can form the belief "I want to take the bus to the airport" because I am aware of my desire of going to the airport and of my belief that the bus will take me to the airport. Forming a belief on basis of E implies being aware that one is in E. It would be impossible for a subject to form a belief on the basis of a state if the subject is unaware of that state. He states that it is sufficient if one just "experiences" the states in which one is, without forming any conscious belief (2008, p. 531).

It could also be thought that the since Fernández's formulation only considers beliefs, it would be unfair to claim that his proposal does not provide self-knowledge. However, this

is not the case, for he explicitly claims that he is proposing a general account of self-knowledge (2008, p. 519). He claims that his method produces self-knowledge given the qualities of asymmetry and strength. Asymmetry comes from the claim that, when applying the bypass view, one forms a belief that is not based on reasoning nor behavioral evidence, which are the only ways to attribute mental states to others. Strength, on the other hand, comes from the claim that the transparency method is invulnerable to the type of errors that arise from ascribing mental states through reasoning or behavioral evidence, e.g. making an incorrect inference or perceptual errors.

In the remaining part of this section will argue that, at least in this formulation, it is difficult to consider the bypass view a theory of transparency in the way that we have presented it before.

Let us first consider instrumental desires. As we have seen in the example of the bus to the airport before, the belief "I desire *p*" is grounded on a belief-desire pair of the form "I desire *q*" and "If *q* is the case, then *p* will be the case". To know this first instrumental desire, we have to know the belief-desire pair that grounds it. In our case, I only know that I desire to take the bus because I know I desire to go to the airport, and because I believe that if I take the bus it will take me to the airport. This is problematic for a transparency theory because the question is mentally directed. In other words, in order to know my desire, first I need to know my beliefs and desires. We seem to fall into circularity here, since the only way to know my instrumental desires is to know my desires. This, however, could be solved if the question regarding the grounding of non-instrumental desires is a world-directed question.

Now we will consider non-instrumental desires, and we will see if the method applies to these is more promising. According to Fernández, we know some of our desires because they are grounded in urges. Someone is justified to believe that she desires to eat ice cream, for example, when that person has the urge to eat ice cream, and the desire is based in that urge. What Fernández seems to be saying here is that the difference between an urge and a desire is that a desire has propositional content, while the urge is pre-propositional (or pre-conceptual). The urge would be more similar to a feeling, or something like that, while the desire is a propositional attitude about that feeling.

This picture presents two main problems. The first one is that the distinction between urges and desires is not clear. Fernández does not give any reason for accepting this

distinction, and it goes against the basic intuition that urges are the most basic type of desires. We tend to think that having an urge for eating something sweet, taking a walk, or singing a song are paradigmatic examples of desires, regardless of our ability to conceptualize them or putting them in propositional form. It is far from clear why do we have to accept this distinction besides that it is a necessary distinction for the bypass view. If we consider again the case of the child with sexual urges who lacks the concepts to understand his own desires, it would be quite strange to say that the child does not have any desire just because he lacks understanding of the object of his desire. Instead, it would seem more natural to say that he does not know how to *satisfy* his desire, or that it is unclear to him what is the exact object of his desire. The example does not work unless we restrict desires to propositional attitudes that necessarily possess conceptual content that can be rationalized by the subject. Fernández, however, does not propose any reason for believing that this is actually the case, and it seems that it is a very strong view about the nature of desires.

Furthermore, even if we accept the distinction between desires and urges, the bypass view still lacks the qualities that are necessary to be considered an instance of the transparent method. If we know that I desire to leave the party in which I am, it is because I know that I have an urge to leave the party. How do I know about that urge? Maybe I know about it because I feel it (Fernández usually says that one *experiences* these types of states). But it cannot be a physical sensation. There could be a physical unpleasant sensation of anxiety, for example, but this physical sensation is not enough to explain the urge of doing the action of leaving the place in which I am. If I can know about my urges, and if urges are "groundless", as Fernandez seems to think, then I cannot know my own urges through the transparency method.

Ashwell (2009) points out that Fernández's account is obliged to postulate some type of inner sense or inner scan that would allow us to have epistemic access to such urges. I agree with her criticism, and it can be strengthened with the following example. To answer the question "Do I desire to leave the party?", I would have to answer the question "Do I have the urge of leaving the party?". This, again, is doubly problematic. First, because it is, as in the case of non-instrumental desires, a mind-directed question. The urge cannot be *in* the object of my desire; it must be part of the attitude I have towards that object. Therefore, I have access to my desires by having access to my mind. Second, and maybe more

importantly, the knowledge of our desires is based on the knowledge of our urges and the knowledge of our urges does not seem to be transparent. Therefore, according to this account, the transparency method has to rely on a non-transparent method to be appropriately applied. If this is true, the metaphysical economy of the transparency method would fall apart, and we would have to postulate again a mysterious inner sense that could account for the knowledge of our urges.

In response, Fernández could respond to this by saying that we do not need knowledge of our urges, in a strict sense of the word. To follow one of his examples, we could say that I know my desire to drink water on the basis of *feeling* thirst. This, let's suppose, might be the case. I know my desire because I have a physical unpleasant sensation, and this physical sensation is the basis of the knowledge of my desire. But let's think about the case of the party again. I might have an unpleasant sensation, but that is not enough to know that I have a desire to leave the party. It is unclear what the urge would be in this case, for it does not seem that it can be reduced to a physical sensation. If I were just to feel an unpleasant sensation in this situation, there would be a variety of possible states of affairs to deal with that situation. I could, for example, leave the party, or go and lock myself in the bathroom. I could even force myself to go to the dance floor and dance, for I could believe that this would alleviate this sensation. In one word, there is no clear, direct relation between a physical sensation and the knowledge of a more complicated desire such as "I want to leave the party", for the physical sensation could ground several possible desires. Maybe Fernández could argue that urges are more than physical sensations, but it is unclear what they could be besides this.

Let's consider briefly the case of value. A similar line of thought can be applied in the case of value. If I value p, under normal circumstances, then I can justifiably believe that I desire p. But how do I know if I value something or not? It is difficult to see how I can answer this question with a world-directed question, since the value is something that I give to the object. The subject is the one that values, and this action is the one that gives value to the object or the action. The question about values and desires, according to Fernández's account, would also be a mind-directed question. Therefore, this account would fail to satisfy the second criterion, for it is unable to provide a world-directed question to account for our self-knowledge of desires.

In conclusion, while the bypass view states that we should "look past our desires" to get to know them (i.e. we have to look at the state that grounds them), the general problem it faces is that "looking past" or "looking beyond" is not directed to the world, i.e. the intentional object of the desire, but to other mental states, such as urges and values. This means that the bypass view does not provide world-directed questions that would serve as a transparent method for knowing our desires. Besides this point, other mental states, such as urges, do not seem to be transparent by means of world-directed questions. If this is the case, a defender of the bypass view is forced to postulate some type of inner sense or inner scan to account for the privileged access that we have of our own urges. To sum up, the bypass view is unable to provide a broad transparent method to account for the self-knowledge of desires.

3. The desirability view

Next I will consider the most influential proposal of the transparency method applied to desires, formulated Alex Byrne (2018), a leading advocate of the Uniformity Thesis[6]. Byrne's proposal comes from an attempt to find a general rule that would allow us to apply the transparency method to all desires. His proposal is based on the idea that desires are intimately connected to the concept of *desirability*, understood as "the qualities which cause a thing to be desired: Pleasant, delectable, choice, excellent, goodly" (2008, p. 160). When these two ideas are applied, we get the general rule DES:

DES If ϕ-ing is a desirable option, believe that you want to ϕ (Byrne, 2018, p. 161).

[6] To defend the uniformity of self-knowledge, Byrne gives an argument based of Shoemaker's idea of self-blindness (1994). According to Shoemaker, a self-blind individual would be a person that lacks every type of introspective access to the phenomena that occur in her mind. This individual, however, would be able to gain knowledge of her own mental states through the third-person means that other persons have to gain knowledge of her mental states, such as her behavior. According to Shoemaker, self-blindness is impossible. Byrne uses this idea to argue that for the generality of transparency. The argument is the following. If transparency applies partially, we would also need some type of inner sense to account for some propositional attitudes. A case of self-blindness would cause dissociation between some propositional attitudes and others, since this individual would be able to know some of her mental states through the transparency method and others only by third-person means. This dissociation, however, is nowhere to be seen. Furthermore, self-blindness does not seem possible. Therefore, transparency has to apply generally.

If I consider, for example, going to the beach and reading as desirable options among the things that I can do (because I think that would be a pleasant experience), then I can *safely* believe that I have the desire of going to the beach to read.

Byrne defends DES with three basic claims: 1) DES is not self-verifying, it is only *practically* self-verifying; 2) DES is not circular; 3) DES is defeasible.

The first claim is understood better when we compare Byrne's formulation of the transparency method for desires with his transparency method for beliefs. Consider Byrne's rule for belief:

BEL If p, believe that you believe that p (Byrne, 2018, p. 103)

BEL is strongly self-verifying. This means, according to Byrne, that "Following BEL *guarantees* (near enough) that one's belief about one's beliefs are true" (2018, p. 115). This self-verification is explained by the fact that "recognizing that p is […] coming to believe that p" (*ibid.*). If I, for example, apply the rule, and p in this case is the proposition "The pigeon is standing on the bench", the truth of "I believe that p" would be guaranteed by the fact that I judged that p is the case, and it is this judgment that makes me a believer of p. It is *strongly* self-verifying because the truth of the second-order belief is guaranteed even if I just try to follow the rule, i.e. if I judge that p but p is not the case. The intention behind this notion of self-verification is to provide a rule that gives a strong justification to the beliefs of our own mental states. Knowledge, in this sense, implies that the second-order belief that results from following the rules must be true.[7]

BEL is, however, different from DES. There are clear cases in which I could consider something desirable and not desire it, which implies that following DES might lead to a false second-order belief. Consider Byrne's example. It is a sunny morning. I am lying on the couch. The weather is perfect. I know that a bike ride by the sea would be a desirable thing to do. That is, it would be pleasant and enjoyable. However, if I come to the conclusion, following DES, that I have the desire of taking a bike ride by the sea I would be wrong, for I

[7] Byrne uses several epistemic terms to characterize BEL. He claims that BEL guarantees truth, that it is reliable, that it produces knowledge, and that following BEL produces safe beliefs (using the Williamson's and Sosa's terminology) (2008, pp. 105-106). I will focus, mostly, on safety and self-verification.

am laying on the couch "wallowing in my own misery" (p. 161), without any desire to go outside[8]. For this reason, DES is not formulated as a strong logical implication of the form "If ϕ-ing is a desirable option, you want to ϕ". Instead, DES is formulated from a more normative stance that would allow us, under normal circumstances, to have a reliable belief (it could not have easily been false).

This is important if we recall the infallibility thesis. Byrne affirms that DES is practically self-verifiable. This means that in almost every case following the rule would end up in a true belief. Besides, the belief that comes as a result of following DES is also a *safe* belief, meaning that it is a belief that "could not easily have been false" (p. 130). According to Byrne, these two elements –a rule that is practically self-verifying and the fact that it produces safe beliefs– are enough to claim that DES produces knowledge while allowing for defeasibility.

Byrne also claims that DES is not circular. Therefore, it is not trivially true. I will not enter in the discussion of circularity here, but the basic claim of the circularity argument against DES is that something is desirable precisely because it is desired. If this were the case, then DES would always be true, but it could not produce knowledge, for we would only repeat what we already knew. This is the type of argument that Andreotta (2020) uses against Byrne, as we will see in section 4.

The third claim is that DES is defeasible. This is closely related to the claim that DES is practically self-verifying. The point of defeasibility is that there could be external factors that block the inference of DES, making the inference invalid. According to Byrne, a clear example of this would be a case in which intentions clash with desires. In other words, a case in which one intends to do an action that is incompatible with a desirable option. Coming back to the example of the bicycle, I do not desire to go cycling by the sea because I intend to stay laying on the couch, even if staying on the couch is less desirable than to go cycling (or even if it is not desirable at all). This would make DES compatible with errors of self-ascription of mental states. Similar cases, however, do not affect, according to Byrne, the safety claim, for he argues that these cases of external factors blocking the inference are "atypical" (2018, p. 161).

[8] One clear indicator of my lack of cycling desires is that I am still laying on the couch without moving. At least this is a third-person view indicator that must be considered.

Now that we have a clear understanding of the desirability view, I will put forward two arguments against this account. The first problem has to do with the direction of the relevant questions provided by the method. Let's call this the direction problem. The second problem is that Byrne's view cannot account for the notion of epistemic asymmetry, which is essential for transparency views. Let's call this the asymmetry problem. As I will show later, both of these problems are related to the notion of desirability.

Let's consider first the direction problem. As stated before, this view presupposes that something is desirable because it possesses some compound of qualities that make it desirable. But what are exactly these qualities? When I consider the bike ride by the sea there is *something* in this state of affairs that makes it desirable. The most natural reading is that I would feel some type of pleasure and well-being if this state of affairs becomes actual. The question "Is going for a bike ride by the sea desirable?" would be a question about the world, for a possible state of affairs in the world. I could say, for example, that it is known that going for a bike ride will improve my health, both physically and mentally, or that it is known that feeling the sea breeze of a sunny day is, in most cases, a pleasurable sensation. These are all world-directed considerations, for I am considering a possible state of affairs and not my mental states.

Roche (2023) points out that there are cases in which desirability becomes more subjective, more subject-directed, causing what is desirable to me to not be desirable to someone else. Let's say, for example, that I consider playing chess desirable because I find mental calculation pleasurable. Another person, however, might find chess terribly boring, and that person might prefer to spend her time playing videogames. This seems like a desirable outcome for Byrne's view, since it seems that it is an empirical fact that different people desire different things.

If the notion of desirability is more subjective, in the sense just sketched, then a person could take into consideration her own beliefs and previous desires to decide whether something is desirable or not. If I believe that playing videogames helps to control my anxiety, for example, then I would take this belief into consideration when considering if playing videogames is desirable. The problem with this conclusion is that the desirability of a thing or a state of affairs would not lie only in the qualities possessed by that thing or state of affairs, but also on the beliefs and attitudes that I have towards that state of affairs. The

seeming world-directed question given in DES would be based on mind-directed questions. If this is the case, then we do not have a true transparency method in DES.

Roche (2023) gives three answers to this problem. The first one is that one can always make desirability judgments from a purely world-directed point of view, and to consider one's own mind is completely optional. The second is that taking into consideration one's own mind does not make the content of the judgment mentalistic. Lastly, the third is that the beliefs that might be taken into account to produce desirability judgments would also be a product of the transparency method. I will now consider these answers in order.

The first consideration is problematic because it seems to deny the subjective view about desirability that we accepted earlier, and this blurs the difference between what I consider desirable and what other people consider desirable. If my beliefs are not necessary to form any judgment about desirability, then it seems that everyone could, at least in theory, make the same judgments about what is desirable or not. Besides, this answer leaves unanswered two important matters: why do we consider our beliefs to judge the desirability of a state of affairs? And, secondly, am I making a world-directed question if to take into account my own beliefs?

The second answer of Roche seems to be a response to this last worry. To defend his claim, he provides the following example: I judge that a bridge is unsafe, partly because I feel fear when I cross it (Roche, 2023, p. 14). According to Roche, taking my fear into consideration does not make the content of the judgment mentalistic. After all, the judgment is about the bridge, not about my own mind. Being unsafe is a quality of the bridge that does not depends on my beliefs or feelings about it. This would be analogous to saying that being desirable is a quality of the object; the reasons that one has for attributing that property might be subjective, but that does not make it mind-directed. The problem is that this could end up in a formulation of the form "The bridge is unsafe *for me*" (not meaning that you are more prone to accident on the bridge than others, but that it is you that judges the bridge to be unsafe). If we reformulate DES in these terms we would get: if ϕ-ing is a desirable option *for you*, believe that you want to ϕ. What is desirable for you could be not desirable for others. This means that being desirable is not a property of the object or state of affairs, but it is given by a subjective judgment. Maybe Byrne's account could accept this without much harm, but then we could not say that desirability is a quality of the objects we desire.

Lastly, if we were to apply DES considering beliefs gained through BEL, and if we assume that BEL poses a pure world-directed question that does not rely on the knowledge of any other mental state, then we wouldn't be, strictly speaking, applying the same method. We would have a transparent method for beliefs that does not take into consideration other beliefs, and a partially transparent method for desires that takes into consideration transparent beliefs. These cannot be the same method, since a fundamental criterion for considering a method an instance of the transparency method is the direction of the question it poses.

Maybe one could think that these problems could be sidestepped if we abandoned the subjective interpretation of desirability and relied on a more "objective" notion. By this I mean that what is desirable is strictly speaking some quality of the object or state of affairs, which would mean that everyone, or at least similar people, should arrive at the same desirability judgments. The problem with this proposal is that we would lose the epistemic privilege that the transparency method attempts to explain. More specifically, we would not be able to account for the epistemic asymmetry that exists between the first and the third-person perspective. Consider the following rule proposed by Doyle (2018):

FDES: If Φ-ing is a desirable option, believe that Alex wants to Φ. (2008, p. 3)

Doyle uses this rule to show that the notion of objective desirability can provide a rule that is not self-verifiable, defeasible, and that produces safe beliefs (in the sense considered above) about other people's mental states. If this were the case, then it seems that FDES also produces knowledge, just as DES does, but not about my own desires, but about Alex's desires. If I use the same method as Alex to get to know her mental states, then it cannot be true that Alex possesses some first-person warrant for the beliefs about her desires that I, an external observer, do not possess. If we use this interpretation of desirability, we seem to lose the first-person perspective that characterizes the transparency method, making it impossible to defend the existence of privileged access.

4. The conceptual approach

Andreotta (2020) rejects both the bypass method and DES. He rejects the bypass method because, according to him, Fernández does not provide a general rule or a general world-directed question that can be applied to every case of desire. The notion of states grounding desire is too broad for a clear application of the transparency method. On his account, an acceptable transparency method must provide a world-directed question that could be applied generally.

On the other hand, he rejects DES because of its circularity. Something is desirable, or doing something is desirable, he argues, exactly because the subject desires it. Being desirable and being desired, according to him, end up being the same thing. If this were the case, then the rule would be completely useless as a method for knowing our own desires. Byrne acknowledges this worry and provides some considerations against the circularity claim. I will not, however, consider this issue here.

After rejecting both proposals, Andreotta proposes his own method, which he calls the *conceptual approach.* His argumentation has three parts. First, he claims that an actual transparent method must use conceptual entailment. Second, it must provide a general world-directed question (as Byrne does). Third, the result of the method must be supported by Moore-paradoxical sentences. Let's consider the three points in more detail.

Andreotta's view is stronger that Byrne's, for he uses the notion of conceptual entailment (Byrne, on the other hand, is satisfied by just talking about inference). The introduction of this concept is quite obscure. When talking about the transparency method applied to beliefs, he states that judging that p does not involve a causal relation with believing that p, but that the judgment *conceptually* entails the belief. To explain his idea, he uses the following example. When one loses one's temper, one might shout. The loss of temper does not cause the shouting, but it conceptually entails it. The point, as I see it, is that the action of believing is conceptually tied to the action of judging, just as losing one's temper is conceptually tied to shouting. In one sense, then, believing that p is coming to judge that p; therefore, one cannot deny p when one judges that p. There is a conceptual link between judgment and believing that is necessary for the transparency method to work.

While trying to apply the notion of conceptual entailment to desire, he states that we must understand the concept of desire. Desires, according to this view, are "a type of mental state that typically instils a pleasurable experience in a subject who imagines a certain state

of affairs occurring" (Andreotta, 2020, p. 200). In other words, desiring is a mix of imagining a state of affairs and some type of pleasurable experience that is, somehow, caused by this act of imagination.

From this notion of desire, we can extract a world-directed question that would be applicable for a transparency method applied to desires. The formulation of the conceptual approach to desires is the following one:

> **The Conceptual Approach to Desire:**
> The question 'Do I desire to ϕ?' (or 'Do I desire that P?') is transparent to the question 'Would ϕ-ing bring me pleasure or satisfaction?' (or 'Would P's occurrence bring me pleasure or satisfaction?'). (Andreotta, 2020, p. 201)

If my friend, for example, calls me and asks me "Do you want to go to the park?" I could apply the conceptual approach and ask myself "Would going to the park bring me pleasure or satisfaction?" Let's say that I imagine the situation, and judge that it will bring me pleasure. I realize that having a chat with my good friend while walking through the trees would bring me a sensation of pleasure and well-being. In this case, I would be justified to self-ascribe the desire of going to the park, as simple as that.

To further strengthen his claim, Andreotta, following Moran (2001) and Shoemaker (1996), relates the conceptual approach with Moore-paradoxical sentences. The standard form of Moore-paradoxical sentences is "P, but I do not believe that P". For example, if I say "The train is late, but I do not believe that the train is late" whoever hears me uttering this sentence would think that I am speaking complete nonsense. What makes this type of sentences paradoxical is that we feel that they are completely absurd, and yet they are not contradictory from a logical point of view.

Explaining why these sentences are absurd is beyond the aim of this paper. Andreotta, however, thinks that Moore's paradox supports the transparency thesis. The fact that we find these type of sentences completely absurd, Andreotta argues, shows that there is a deep conceptual relation between judging that p and believing that p, and this relation makes it the case that the question "Do I believe that p" is transparent to the question of whether p.

Andreotta claims that Moore-Paradoxical sentences exist in the case of desires.

Following everything we have said so far, this type of sentence would have the following form: "I think that ϕ would be pleasurable, but I do not desire to ϕ". This sentence, according to this view, is not necessarily false, at least from a logical point of view, but it must consider it absurd. If someone were to say something like this, Andreotta argues, we would say that this person is confused about the meaning of desire. This would show a deep, conceptual relation between pleasure and desire, which connects the judgment about the pleasure that a state of affairs would bring with the desire that this state of affairs happens.

This is, broadly speaking, Andreotta's proposal. I think that we should reject it for three main reasons. First, the idea of conceptual entailment is far too obscure. Second, the objection that Andreotta uses against Byrne can also be applied to his account. Third, the Moore-paradoxical sentences that he proposes for desires are different from the standard sentences that give raise Moore's paradox, and they fail to justify transparency with the same strength.

Let's go back to conceptual entailment. In several instances, Andreotta says that judging something to be pleasurable underpins a desire for that something. In other words, the relation between pleasure and desire would be that pleasure supports, in some sense, desire. What does this mean? It is unclear how to interpret Andreotta's idea. It is clear that it is not a causal relation, for we must be able to account for cases of error (Andreotta claims that his proposal does not defend infallibility). It cannot be, either, that pleasure usually accompanies desire, since Andreotta's claim is much stronger, for he is talking about conceptual *entailment*.

One charitable possibility is that the property of being pleasant would *explain* why we desire something to be the case and not another thing. This would be an acceptable interpretation for the shouting example, as far as we can explain the shouting of someone by saying that that person lost his temper. This interpretation, however, does not work for the case of desire. An explanatory relation is far too weak for the type of epistemic link that we are looking for here. I do not want to explain why I would desire to eat chocolate cake; instead, I want to know if I actually desire eating chocolate cake.

It would also be problematic if we made the connection too strong and said that the *meaning* of desiring something is to find it pleasurable (which is also sometimes suggested by Andreotta). If this were the case, the conceptual approach would be completely circular

and analytic, just as saying that the question "Is this shape a triangle?" would be transparent to the question "Does this shape have three sides, and the sum of its internal angles sum up to 180º?".

One could adventure other interpretation, but this would just strengthen my point: the idea of *conceptual entailment* is far too vague to give an acceptable account of the epistemic link that we are trying to establish.

Let's turn our attention to Moore's Paradox now. I tend to agree with Andreotta when he states that Moore's Paradox can be considered a reason in favor of the transparency thesis. It seems that it shows an actual connection between our judgments about how things are and our beliefs. However, it is not clear whether the paradox can be applied to the case of desire.[9]

If one hears someone saying, "The tree is big, but I do not believe that the tree is big", one would be justified in thinking "this person is raving", and in saying something like "What are you saying? Are you alright?". In other words, you would be justified in dismissing the assertion as blatantly absurd. On the other hand, if someone says, "I think that going swimming would be pleasurable, but I do not want to go swimming", you might find the utterance weird, but you would not consider it nonsense. This is clearly the case because you would be entitled to ask, "How come that you don't want to go swimming, then?", and it would make sense in the conversation. That person might reply that she swims regularly, and most of the time she feels pleasure while exercising in this way, and that she believes that she would feel the same if she were to go swimming, but that she had a very rough week that she is tired and sad and that the only thing that she wants to do is stay home and watch movies. The continuation of a similar conversation is impossible in the case of belief, since in this case no explanation would suffice. The fact that we react differently in conversation to the paradox in terms of desire shows that it does not seem paradoxical for us, at least not in the way intended. We could ask for an explanation, but this also shows that we do not

[9] Boyle (2024) points out that the paradox is best understood in terms of assertion (I assert p and I also assert that I do not believe p). If this is the case, he argues, there are pragmatic considerations about assertion that do not apply when we consider solely the belief. In other words, it seems paradoxical for someone to assert "p" and "I do not believe p" because with "p" she would be pragmatically implying "I believe that p". This would explain the feeling of contradiction that emerges when someone states the paradox. However, appealing to pragmatics is not a good way to approach belief independently of the assertion of these beliefs. Pragmatics, in this case, do not apply. Therefore, using Moore's Paradox to account for transparency may need a special solution to the paradox. The solution must point clearly to the relation between belief and knowledge of the belief without recurring to pragmatic explanations.

dismiss the utterance as nonsense.

With these considerations we can conclude that Andreotta's account is not adequate as a general theory of the transparency method. I think that many of the arguments used here could also be applied to the other intentional attitudes that Andreotta tries to account for (intentions and wishes). I do think, however, that there is some type of fundamental relation between desire, pleasure and satisfaction, which suggests that the conceptual approach captures a key aspect of the transparency method after all. Nevertheless, we need to account for it in different terms, as we will see next.

5. A partial solution

So far I have tried to show that none of the present theories of the transparency method can be applied successfully to the epistemology of desire. Does this mean that we should abandon completely the attempt to use the transparency method in the case of desires? I think not. In this last section of the paper I will sketch a possible solution to the transparency of desires. This solution, however, implies the negation of the uniformity thesis, and establishes transparency as one of the methods that we can use to get to know our desires.

My proposal arises from two elements: 1) the distinction made by Boyle (2009) between active self-knowledge and passive self-knowledge and 2) an understanding of why transparency seems so strong in the case of beliefs.

If we go back to Moran (2001), we find that the connection between judging that p and believing that p is founded upon rational agency. To put it in simple terms, the basic idea is that when we use the transparency method with our beliefs, and we ask, "Is p the case?", we start a process of rational deliberation that ends once we reach some solution and we "make up our minds". According to this account, judging that p is coming to believe that p through a process of deliberation grounded on rational agency. If someone were to ask me "Is Daniel coming to the party?" I might realize that I am not sure. Daniel sometimes goes to parties; sometimes he prefers to stay home. However, I remember that the girl that he likes is going to the party, and that he takes every chance he gets to talk with her. With all these considerations I make up my mind and conclude: "He is coming to the party". In this process,

according to Moran's account, I have come to believe that Daniel is coming to the party, a belief that I did not have prior to my deliberation about all the previous external reasons.

Boyle (2009) argues that the process of rational deliberation can only be applied to a subgroup of our mental states, namely mental states that are formed through deliberation. We also have passive knowledge of the mental states that we form independently from any kind of rational deliberation and that are indeed insensitive to the process of deliberation. Boyle argues that we have active knowledge of our judgments (which would account for the transparency method regarding beliefs) and that we have passive knowledge of our perceptions.

I propose that we should look at the epistemology of our own desires as a combination of both active and passive self-knowledge. Some desires can arise through deliberation, while others cannot. Desires that are actively known are accessible through the transparency method, but a different approach is needed for those that can only be passively known.

Consider the following case. Clara is very hungry. She steps into a place in which there are several restaurants. She does not know, however, which food she would like to get. She sees the hamburger stand, the Mexican stand, the hot god stand, and, when she looks at the pizza place, she comes to the realization that she wants pizza. What is this "realization"? I argue that in this example Clara is not trying to find a desire that was already hidden somewhere in her mind. Instead, I think that this would qualify as an instance of self-determination. In Moran's words, the subject, by considering the different options that are in front of her, her beliefs and, maybe, other desires, "makes up her mind" about what she wants. Prior to this process she did not have any particular desire for a specific type of food, she just felt hungry. This case would be analogous to the party case considered earlier. In both cases the subject is making up her mind. In one case, is about beliefs, while in the other case is about desires.

Someone might respond to this by saying that this is also a case of passive desire. Clara did not choose to desire the pizza. The desire, so to speak, overtook her in the exact moment that she saw the pizza place. I argue that even if this were the case and she could not have desired anything else, there is still a process of deliberation. Clara had in her mind all the other options possible, and preferring pizza in this context implies a comparison with all the other options that she considered before. Even if the desire "overtook" her, it does so as

the result of the comparison between the different options.

On the other hand, if Clara were to wake up from a nap on a Sunday afternoon, and the first thought that popped in her head was "I want some pizza", it is very hard to argue that this is a case of making up her mind. There is no rational deliberative process, which would make this desire impossible to know using Moran's method.

Let's suppose that Boyle's distinction is right, and that there is active and passive self-knowledge. Let's also assume, for the sake of the argument, that we can make up our minds regarding some desires, but not others. Then we can see that some desires have a quality that make them transparent to world-directed questions, while other desires are completely opaque to these same questions. The transparency method could be thus applied to desires, but only in this limited way. The conclusion that would follow is that we cannot generalize the transparency method to every mental state that we have. We must propose different methods for those states that it cannot adequately address.

What account of transparency should we use to know our active desires, then? It is hard to say, for we have seen that the three main theories present different problems. However, we must consider the possibility of reformulating these theories in such a way that the can account for active desires. Consider the following rule:

ADES: If ϕ-ing is considered a desirable option after a process of rational deliberation, believe that you want to ϕ[10].

The main problem with this formulation is that it seems to presuppose that we are able to distinguish between passive and active desires. If the method is supposed to be reliable, then it should be applied only to cases in which the subject is in a position to make up her mind. Imagine that you are choosing between two movies to watch. You consider your options carefully, and you decide that *Star Wars* is the most desirable option. However, before your deliberation, the desire to watch *The Lord of the Rings* emerged in you. You

[10] It is more difficult to reformulate the conceptual approach. One could formulate as a meta-rule that the conceptual approach is valid only when the judgment of pleasure and satisfaction is the result of a deliberative process. But Andreotta claims that imagination plays an important role in the creation of the judgment. Could imagination be part of a rational deliberative process? Intuitively, yes. A more careful discussion, however, is required here.

apply ADES, and the belief that it generates is false because your desire was insensitive to your deliberation. You want to see the *Lord of the Rings*, even when the result of your deliberation says *Star Wars*. Someone could argue, however, that this is a strange case, and that deliberation would allow us to discover our passive desires (in a normal case, my deliberation would end up in the belief that I want to watch *The Lord of the Rings*). If examples like this could be explained as defeasibility cases (as Byrne does) the proposal might be plausible.[11]

6. Conclusion

In this paper I have rejected three different proposals for applying the transparency method to desire: the bypass view, the desirability view, and the conceptual approach. I have shown that each one of them presents problems that make them unable to provide a satisfactory version of the transparency method. I have introduced the distinction between active and passive self-knowledge and argued that integrating this distinction into the debate would be beneficial. Accepting this distinction, however, would imply the rejection of the Uniformity Thesis. This rejection is based on the presupposition that the transparency method actually provides self-knowledge in the case of beliefs. Whether we can formulate a world-directed question that adequately captures the passive-active distinction and thus accounts for the transparency of (some) desires remains an open question.

References

Armstrong, D. M. (1968). *A Materialist Theory of Mind*. Routledge.

Ashwell, L. (2013). Deep, dark… or transparent? Knowing our desires. *Philosophical Studies, 165*(1): 245-256.

[11] I am not claiming that ADES is the correct approach. After all, the objections made against DES apply to ADES. A new formulation is needed. I use it as an example of a possible world-directed question.

Ashwell, L. (2009). *Desires and Dispositions* [Doctoral Thesis, Massachusetts Institute of Technology]. MIT Libraries. https://dspace.mit.edu/handle/1721.1/55176.

Barz, W. (2015). Transparent introspection of wishes. *Philosophical Studies, 172*: 1993–2023.

Boyle, M. (2009). Two kinds of self-knowledge. *Philosophy & Phenomenological Research, 78*: 133–164.

Boyle, M. (2024). *Transparency and reflection*. Oxford University Press.

Byrne, A. (2018). *Transparency and self-knowledge*. Oxford University Press.

Doyle, C. (2019) [Review of the Book *Transparence and self-knowledge* by A. Byrne]. *European Journal of Philosophy, 27*(2): 515-518.

Evans, G. (1982). *The varieties of reference*. Oxford University Press.

Fernández, J. (2003) Privileged access naturalized. *Philosophical Quarterly, 53*. 352–372.

Fernández, J. (2007). Desire and self-knowledge. *Australasian Journal of Philosophy 85* (4): 517 – 536.

Gerter, B. (2003). Philosophical Issues about Self-Knowledge. In Gertler, B. (Ed.), *Privileged Access: Philosophical Accounts of Self-Knowledge* (pp. 19-53). Routledge.

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

Nichols, S. & Stitch, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds.* Oxford University Press.

Paul, S. K. (2012). How we know what we intend. *Philosophical Studies, 161*: 327–346.

Roche, M. (2023). Introspection, Transparency, and Desire. *Journal of Consciousness Studies, 30*(3), 132–154.

Shoemaker, S. (1994). Self-Knowledge and "Inner Sense": Lecture II: The Broad Perceptual Model. *Philosophy and Phenomenological Research, 54*(2), 271–290.

Silins, N. (2013). Introspection and Inference. *Philosophical Studies, 163*(2): 291-315.

Valaris, M. (2014). Self-Knowledge and the Phenomenological Transparency of Belief. *Philosophers' Imprint, 14*(8): 1-17.