In a Deeper Sense[1]

Joshua Knobe
Yale University

[Unpublished manuscript. Comments and suggestions welcome.]

Research on dual character concepts has explored cases in which people think that a term applies to an object in a superficial sense but does not apply to that same object in a deeper sense. Most of this research has focused on cases of one particular type, namely, cases in which the object fails to embody the characteristic values of a particular category. However, there are also other types of cases in which we would be inclined to say that a term does not apply in a deeper sense. For example, we might also say this if an object has acquired a certain status but it failed to acquire this status through a morally or legally legitimate process. Or we might say it if a person has certain emotions but these emotions are not rooted in the person's true self. I argue that these apparently unrelated types of cases are surprisingly similar in a number of important respects. Thus, we should not be seeking a separate account for each separate type of case; we should be seeking an account that explains in a more general way what people mean when they say that a term does not apply in a deeper sense.

Imagine a person who has a job as a biology professor. She spends her days conducting experiments, running statistical analyses, writing up papers. But she is not in any way trying to find the truth. Her only goal is to achieve professional success, and all her work consists of trying to support theories that are completely false but that are being pushed by wealthy funders from industry. Is this person a scientist? Well, there's a sense in which she is obviously a scientist, but at the same time, one might think that there is a deeper sense in which she is not truly a scientist at all. One might say that she is not a *true scientist*.

Now consider a different sort of case. In 1399, when Richard II was ruling as king of England, the man who would become Henry IV launched a successful rebellion. He had Richard II thrown in prison and took over the throne. After these events were completed, Henry IV was ruling as king, while Richard II was languishing in a prison cell. But who was truly the King of England? In a case like this, one might not know quite what to say. There is a sense in which Henry IV was obviously king, but one might also think that there is a deeper sense in which Henry IV was not the king. One might say that he was not the *true king*.

Just looking at these two cases, it might initially seem that they are completely different. In the case of the scientist, it seems that what is going on is that the person displays various superficial features we associate with scientists but that she fails to embody the deeper values that being a scientist is really all about. But none of that is happening in the case of the king. The problem with Henry IV was not that he failed to embody the deeper values that being a king is really all about. (On the contrary, many people think that he embodied the relevant values more fully than Richard II ever

[1] For powerful objections to my previous account of these phenomena, I am grateful to Phillips and Plunkett (2023), Baumgartner (2024), Prinzing and Fredrickson (2023) and Del Pinal and Reuter (2017). The present paper offers a very different account that was developed partially in response to these objections.

did.) Rather, the problem was something else entirely. It was that he did not rise to the throne through the appropriate morally or legally legitimate process. He was not the son of the previous king.

Let's now consider a third type of case. Sarah usually feels pretty cheerful and upbeat, and she often thinks about how everything has been going great for her. But sometimes, when she is alone in the dark of night, she feels like there is a voice coming from deep within her telling her that she is going wrong in a fundamental way and that she ought to be doing something completely different with her life. Is Sarah happy? In response to this question, you might well have the same reaction you did to the previous two. You might think that there is a sense in which Sarah is clearly happy but that in a deeper sense she is not happy at all. She does not have *true happiness*.

This third case seems different from either of the previous two. It is not a matter of whether something arose through a morally or legally legitimate process, nor is it a matter of whether it embodies the right values. Instead, it seems to involve the notion of a deeper level of the self. The thought is that a person can have one emotion at a superficial level of the self but a very different emotion at a deeper level.

Within recent research, there has been a great deal of work on cases like the first we have explored here – cases like the one of our biology professor who fails to embody what science is really all about. This first phenomenon is often referred to as 'dual character,' and a substantial body of work has been devoted to exploring it (Almeida et al., 2023; Baumgartner, 2024; Del Pinal & Reuter, 2017; Guo et al., 2021; Knobe et al., 2013; Leslie, 2015; Liao et al., 2020; Malone, 2023; Neufeld 2022; J. Phillips & Plunkett, 2023; B. Phillips, 2022; Reuter, 2019).

But suppose we now expand our scope and look at all three types of cases. At least initially, it might seem that these three types of cases are completely different from each other. Thus, it might seem that whatever we have learned from existing research about cases of the first type will not also apply to cases of the other two types. What we need here, one might think, is just three completely separate theories, one for each separate type of case.

I will be arguing for the extreme opposite view. I argue that almost everything that is surprising or theoretically interesting about cases of the first type also arises in the same way for cases of the second and third types. Thus, what we really need here is not a theory that applies just to cases of the first type. What we need is a more general understanding of what it means to say that term does not apply to an object in a deeper sense.

2. **Differences between different types of deeper-sense cases**

Before exploring the respects in which these different cases might actually be surprisingly similar, it will be helpful to discuss the ways in which they are obviously very different. In what follows, I will be using the phrase *deeper-sense cases* for cases in which one might think that a term clearly applies to an object in some superficial sense but does not apply in a deeper sense. As we will see, different types of deeper-sense cases appear to be quite different, and indeed one might initially think that they are so different that there is little to be gained by grouping them together.

2.1 *Embodying-values cases*

First, consider cases like that of the biology professor with which we began. In cases of this type, what does it mean to say that a term does not apply in a deeper sense? Within existing research, there has been a large amount of research on this question, and this research has given us at least some preliminary understanding of what is going on (see, e.g., Knobe, 2022). At the core of most proposals is the notion that there are some categories that are associated with certain characteristic values. We might describe these values as being what the category 'is really all about.' For example, we might describe the characteristic values of being a scientist as 'what being a scientist is really all about,' or we might describe the characteristic values of being a Christian as 'what being a Christian is really all about.' In general, the characteristic values of a category will not just be the things that make something a good member of that category. For example, one might think that being knowledgeable and intelligent are things that make someone a good scientist, but if you were asked, what being a scientist was really all about, you would not say that it was about being knowledgeable and intelligent. Instead, you might say that it is about something like: genuinely striving to find the truth regarding empirical questions.

Then what it means to be a member of the category in a deeper sense is that one embodies the characteristic values. Thus, to be a scientist in a deeper sense is to embody the values that characterize the category of scientists. If someone has a job as a biology professor but is not genuinely striving to find the truth, one might think that this person does not embody the characteristic values of being a scientist, and that she is therefore not a scientist in a deeper sense.

This criterion can be applied in numerous different cases. One might think that a person who goes to church every Sunday nonetheless fails to embody what Christianity is all about, and one might therefore say that this person is not a Christian a deeper sense. Or might think that two people who hang out all the time and have lots of fun together nonetheless fail to embody what friendship is really all about, and one might therefore say that they are not friends in a deeper sense. This criterion can also be applied to objects that are not human beings. A painting in an art gallery might fail to embody what art is all about, and one might therefore say that the painting does not count as art in the deeper sense.

In previous work, I mistakenly thought that all cases involving the notion of a deeper sense were best understood in terms of embodying values, and I therefore referred to the phenomenon involved in embodying characteristic values as 'dual character.' This terminology is not ideal. Henceforth, I will refer to cases that specifically involve embodying characteristic values as 'embodying-values cases,' and I will use the phrase 'dual character' for the broader phenomenon that also extends to cases involving other criteria.

2.2. *Legitimacy cases*

Now consider cases like the case of the usurper king. Here again, we might again be drawn to say that a term applies to an object in some superficial sense but not in a deeper sense, but in this second type of case, it seems that something importantly different is happening. The reason we think that the term does not apply in a deeper sense is that we think that an object did not acquire the relevant status through a (morally or legally) legitimate process. When it comes to being king, for example, one might think that one legitimate way to become a king is to be the first-born son of the previous king. Since Henry IV did not become king through a legitimate process, one might think that he was not truly the king in a deeper sense.

Numerous other cases seem to make use of this same criterion. For example, suppose that an indigenous tribe is living on some land, but then a robber baron has them violently removed and begins controlling the land himself. At this point, the robber baron completely controls the land, and there is no chance that the indigenous tribe will ever get it back. Who is the owner of the land? In this case, we may be inclined to think that the robber baron is not the legitimate owner. So we might say that although there is a superficial sense in which he is obviously the owner, there is a deeper sense in which he is not the owner at all.

Or consider the ordinary notion of consent. Suppose that a person does say 'yes' but that this person is a child, or is intoxicated, or has been coerced. Has the person consented? We might think in such a case that this was not legitimate consent. So we might say that although there is a sense in which the person has clearly consented, there is a deeper sense in which she has not consented at all (see, e.g., Demaree-Cotton & Sommers 2022, discussed below).

The key point for present purposes is that legitimacy cases appear to be very different from embodying-values cases. When we say, e.g., that there is a deeper sense in which the robber baron is not the owner of the land, we are not saying that he fails to embody the right sorts of values but rather that he failed to acquire the land through the right sort of process. Thus, it might seem that the claim that there is a deeper sense in which the robber baron is not the owner is fundamentally different from the claim that there is a deeper sense in which the biology professor is not a scientist.

2.3. *True-self cases*

Finally, consider attributions of psychological states. We might say of a certain person that there is clearly a sense in which she is happy but that there is also a deeper sense in which she isn't truly happy at all. Similar claims could be made about love, sadness or hatred. Thus, we can ask whether a person is experiencing 'true happiness,' 'true love,' 'true sadness,' 'true hatred.'

Existing research suggests that these judgments should be understood in terms of the notion of a *true self* (Newman et al., 2015; Prinzing et al., 2023). On this view, the claim that there is a sense in which someone feels an emotion but a deeper sense in which they do feel that same emotion means something like: The person feels an emotion on a superficial level of the self, but does not feel it deep down in her true self. Of course, difficult questions arise about how to understand the distinction between a superficial level of the self and a true self that lies deeper down, but we will not be exploring those questions further here.

Instead, the key point for present purposes is just that this notion is importantly different from either of the two notions we explored previously. Consider a teenager of whom we could say both 'She hates her little sister' and 'In a deeper sense, she does not truly hate her little sister.' This second claim does not mean that she fails to embody the values that hatred is really all about, nor does it mean that she did not acquire her hatred through a morally or legally legitimate process. It means something different from either of those two things: it means that her hatred is not rooted in her true self.

2.4. *Summing up*

This concludes our discussion of the criteria people used to determine whether a term holds of an object in a deeper sense. The principal conclusion has been that people use different criteria in different cases. In some cases, they ask whether the object embodies certain deeper values; in others, they ask whether something arose through a morally or legally legitimate process; and in yet others, they ask whether something is rooted in a person's true self. Presumably, there are also other types of cases.

Although I will be arguing that there are fundamental similarities between these types of cases, I do not mean to be denying the obvious differences. As an analogy, consider the differences between different flavors of modals. We can use modals to make claims about moral rules ('You can't keep treating her like that') and also to make claims about physical laws ('No particle can go faster than the speed of light'). The usual view is that there are very fundamental connections between modals of these different flavors, but this usual view should not be construed as denying the obvious ways in which claims about moral rules are different from claims about physical laws.

3. **Similarities among deeper-sense cases**

In this section, I provide evidence for the claim that there are fundamental similarities between deeper-sense cases of different types and, indeed, that almost everything one might see as surprising or theoretically important about embodying-values cases also arises for legitimate cases and true-self cases.

I do not have a general theory of deeper-sense cases, and in what follows, I will try to be as explicit as possible about aspects of these phenomena that I do not understand. Thus, I will be providing evidence that different types of deeper-sense cases share certain features, but I will not be offering a theory that explains why different types of deeper-sense cases share these features.

Still, even in the absence of a well-articulated theory, we can offer at least a preliminary picture. This picture would be that people have a general notion of what it means to say that a term does not apply to an object 'in a deeper sense.' This general notion is expressed in certain characteristic phrases, and it has certain characteristics downstream consequences. Of course, the notion manifests itself differently in different cases, and in particular, there are differences in what people think the relevant object is lacking (embodying values vs. legitimacy vs. true self). Yet, despite these differences in manifestation, it is possible to develop substantive generalizations that apply across the board to cases in which someone says that a term does not apply in a deeper sense.

3.1. *Linguistic expressions*

Let's begin with the most straightforward piece of evidence. In embodying-values cases, people can add certain extra words or phrases to indicate that they intend a term to be understood in the deeper sense. Strikingly, those exact same linguistic expressions can also be used in legitimacy cases and true-self cases.

For example, numerous studies have explored the ways in which people can use the word 'true' in embodying-values cases to indicate that they intend a term to be understood in the deeper sense (e.g., Bailey et al., 2021; Del Pinal & Reuter, 2017; Knobe et al., 2013). But importantly, that exact same word can be used in all three types of cases.

(1) She is not a true scientist.
(2) He is not the true king.
(3) This is not true happiness.

Indeed, a recent study finds that phrases like 'true happiness' and 'true hatred' are specifically used to refer to emotions that are rooted in the true self (Prinzing et al., 2023).

Similarly, it seems the phrase 'in a deeper sense' can be used in all three types of cases.

(4) In a deeper sense, she is not a scientist.
(5) In a deeper sense, he is not the king.
(6) In a deeper sense, she is not happy.

And 'ultimately' can be used in all three.

(7) Ultimately, she is not a scientist.
(8) Ultimately, he is not the king.
(9) Ultimately, she is not happy.

Now consider what happens when we use 'real' in each type of sentence.

(10) She is not a real scientist.
(11) He is not the real king.
(12) This is not real happiness.

In all three cases, a sentence with 'real' can mean the same thing that the sentence with 'true' does, but it does not necessarily have to have that meaning; it can also be used to mean other things. For example, the sentence 'She is not a true scientist' seems to mean something like: she does not truly embody the values that are characteristic of science. Then the sentence 'She is not a real scientist' can be used to mean that same thing, but it can also be used in other ways. (It could be used to describe a spy who is pretending to be a scientist.)

The similarity in linguistic expressions across these three types of cases is quite striking. In fact, I have not been able to think of any linguistic expressions that can be used to pick out the deeper sense in embodying-values cases but cannot also be used to pick out the deeper sense in the other two types of cases.

Moreover, this does not appear to be an idiosyncratic fact about the English language in particular. Other languages also have individual linguistic expressions that can be used to pick out the deeper sense in embodying-values cases, legitimacy cases, and true-self cases. For example, the Spanish word *verdadero*, the Hebrew *amiti* and the Lithuanian *tikras* can all be used in all three types of cases.

These patterns seem to suggest something fundamental about what these linguistic expressions mean. Given that there is a broad pattern such that the same expressions can often be used in all three cases, it seems that we have at least some reason to reject any theory according to which these expressions have a completely different meaning in each case. For example, we have some reason to reject any theory according to which the word 'true' means three different things in the expressions 'true scientist,' 'true king' and 'true happiness.' Instead, the natural hypothesis would be that there is a single meaning that the word has in all three of these cases, and that this meaning is something like *in a deeper sense*. But, of course, for this hypothesis to be right, it would have to be the case that there is a unified notion of a term holding in a deeper sense that can be applied in all three cases.

3.2. *Ambiguity and confusion*

Thus far, we have been discussing the ways in which people can add extra words to indicate that they intend a term to be understood in a deeper sense. But what happens when people use these terms in sentences that do not also include any of these extra words? Suppose a person just says 'She is a scientist' or 'He is a king' or 'She is happy.' Will the term then be understood in the deeper sense or in the more superficial sense?

Looking across our three types of deeper-sense cases, we seem to have a unified answer to this question. The answer is: People find these sentences confusing. If one doesn't do anything to specify whether one means the superficial sense or the deeper sense, then, in certain kinds of cases, there will be no clear consensus response as to whether the sentence is true or false.

First, consider embodying-values cases. Imagine a person who is always hanging out with you and doing fun things with you. Now suppose that, ultimately, this person does not truly have your back. In such a case, would you or would you not agree with the following sentence?

(13) This person is a friend.

If you are like most people, you will find this case confusing. You might think that (13) is obviously true in one sense but completely false in another sense. Confirming this intuitive point, one study gave participants embody values cases involving a number of different terms, including friend, scientist, artist, etc., and asked in each case about a sentence like (13). Overall, the results did not indicate that participants strongly agreed or strongly disagreed. Instead, mean ratings for these sentences tended to be at approximately the midpoint of the scale.

Now turn to legitimacy cases. In one study (Demaree-Cotton & Sommers, 2022), participants were given cases in which an agent does say 'yes' but in which this agreement does not arise through a morally or legally legitimate process. For example, in one case, participants are told that Frank asks Ellen to have sex. Ellen says 'yes,' but she has an intellectual disability that makes her unable to make decisions about these questions in the usual way. Participants were then asked whether they agreed or disagreed with the sentence:

(14) Ellen's 'yes' didn't count as consent.

In a case like this one, people might think that there is a superficial sense in which the agent has consented but also a deeper sense in which she has not consented. So what do they say about a sentence like (14)? Here again, the answer is that people did not fall clearly on either side.

Finally, consider true-self cases. In one recent study, participants were given a series of vignettes about an agent who has an emotion at a superficial level, but does not have the emotion deep down in his or her true self (Prinzing et al., 2023). They were then asked whether it was right or wrong to say that the agent had that emotion. For example, in one vignette, participants were told that Mario is happy on a superficial level but was not happy deep down in his true self. Then they were asked whether they agreed or disagreed with the sentence:

(15) Mario is happy.

As in each of the other cases, there was no strong consensus in favor of either answer to the question.

The obvious explanation for these findings would be that in most ordinary contexts, people do not find it important to make it clear whether they are using a term in the superficial sense or the deeper sense. They simply assume that the distinction between these senses will not make any difference. Then, in the special case where it happens that a term holds in the superficial sense but not in the deeper sense, it will be genuinely unclear how to apply the term they have used to the object in question.

Consider an ordinary conversation in which a person asks: 'Who was the king of England in 1400?' In asking this question, the person presumably does not specifically mean either (a) king in the superficial sense or (b) king in the deeper sense. So if you think that in 1400 someone was king in the superficial sense but not in the deeper sense, there will be no clear answer to the question as originally posed. You will have to introduce further distinctions.

3.3 *Competing criteria vs. retaining both criteria*

Thus far, we have simply noted that in deeper-sense cases, people seem to have two different criteria for the application of the same term. But of course, there are also plenty of other kinds of terms that have this property, and indeed, there is already an enormous literature on different ways in which this sort of property can arise. Still, even though there are many other terms that have broadly this same property, the property seems to manifest itself very differently in deeper-sense cases.

In most ordinary cases, the different criteria seem to compete with each other. If we choose to use one criterion, then by that very fact, it seems that we are choosing not to use the others. To illustrate, consider a simple dialogue:

A: I grew up in a middle-class household.

B: Middle-class?!? I can't believe you would call the household you grew up in 'middle-class.' You grew up *rich*.

Here, Speaker A is trying to use the term 'middle-class' with certain criteria. Speaker B proposes different criteria and, in doing so, completely rejects the criteria that were originally used by A. Existing research has led to the development of sophisticated frameworks for thinking about cases like these (e.g., Barker, 2002; Khoo, 2020; Plunkett & Sundell, 2013).

Although we will not be able to get into the details of these frameworks here, it will be helpful to introduce one core idea. The usual approach is to model this phenomenon in terms of speakers proposing changes to the conversational context. Suppose we are in a conversational context in which it is unclear which criteria should govern the use of a particular term. If a speaker in this context utters a sentence that would be true on some criteria but not others, she thereby proposes to shift to a context in which the term is governed by criteria that make her sentence come out true. Thus, when B says that A is not middle-class, B proposes to shift to a contest in which the criteria for the 'middle-class' make it right to say that B is not middle-class. And now comes the important point: The speaker thereby also proposes to reject or eliminate the criteria that would make it right to say B is middle class. Thus, if she succeeds in moving the conversational context in the direction she proposes, we will be in a context in which we do not still have the criteria that make it right to say that A is middle class.

But now consider what happens in deeper-sense cases. For example, consider this dialogue:

A: I've been reading more about Ayn Rand's philosophy.
B: Philosophy?!? That stuff? Ultimately, her work isn't even truly philosophy at all.

In this second dialogue, it might initially appear that B is doing basically the same thing she does in the previous dialogue. That is, it might seem that she is trying to introduce one criterion and thereby reject another criterion. But this is not quite right. What is going on in this second dialogue is importantly different and can't just be straightforwardly modeled using the frameworks that have been introduced to understand the first.

The key difference is that when B suggests that there is a deeper sense in which the term does not apply to an object, she is not trying to shift to a context in which we reject or eliminate the superficial sense according to which the term does apply to the object. On the contrary, if you do not understand this superficial sense, you cannot understand what she is trying to say. Her whole point is that although the term *does* apply in a superficial sense, it does not apply in a deeper sense.

Perhaps the best way to see this is to contrast our two dialogues. In the first one, B sees that there is a standard according to which certain people count as middle class, and B is then trying to completely reject that standard. But this is not at all what is happening in the second dialogue. In that dialogue, A is invoking a criterion on which we distinguish works of philosophy from texts of other kinds (novels, scientific papers, restaurant reviews, etc.). Then B is not trying to get rid of that criterion. She is not trying to suggest that Ayn Rand's work should be treated just like any other non-philosophy text. Rather, you can't understand what she is trying to say unless you understand that (a) she recognizes that there is a superficial sense in which this work is obviously philosophy and (b) she is now trying to assert that there is a deeper sense in which it is nonetheless not philosophy.

Importantly, this very same phenomenon also arises for other types of deeper-sense cases. For example, consider a legitimacy case:

A: I recently talked with the king.
B: The king?!? Henry IV has never really been our king. The true king is still Richard II.

Here again, we see a speaker saying something that one can only understand if one keeps in mind two different criteria. The speaker is not trying to say that Henry IV is just like any other person who is not the king. Rather, the whole point is that there is a superficial sense in which Henry IV is obviously the king but also a deeper sense in which he is not truly the king at all.

Finally, consider a true-self case:

> A: Then she fell in love with Desmond.
> B: Love?!? What she feels for him isn't really love at all. She only feels that way about him because she knows he's the one everyone thinks she ought to feel that way about.

The point appears to arise in exactly the same way here as well. The speaker is not trying to say that this person is just like all of the many other people who are not in love with Desmond. Instead, the whole point is that although there is indeed a superficial sense in which she is in love with Desmond, there is also a deeper sense in which she does not truly love him at all.

3.4. *Normative significance*

I have been suggesting that deeper-sense cases are characterized by an awareness that a term applies to an object in one sense but not in another. One striking fact about these cases is that people tend to agree with sentences that explicitly express both of these claims at the same time. For example, people agree in certain cases with the sentence:

> (16) There is a sense in which this is clearly art, but ultimately, you would have to say that there is a deeper sense in which it is not art at all.

Within the existing literature, sentences like this one are referred to as *dual character statements*, and there has been a great deal of research on the conditions under which people think that they can be used (e.g., Knobe et al., 2013; Phillips & Plunkett, 2023).

Of course, people do not often use anything like an explicit dual character statement in ordinary life. The reason they are worth studying is not that anyone would ordinarily utter a sentence like this but rather that they make explicit what people are trying to convey implicitly in more ordinary deeper-sense cases. Someone might point at a painting in an art gallery and say:

> (17) That is not art.

Though this sentence is only four words long, the full dual character statement (16) brings out more explicitly what (17) is trying to convey. That is, in a case like this, the speaker is not trying to deny that there is a sense in which the painting is art; she is just trying to assert that there is a deeper sense in which it is not art.

Within existing research, there has been a lot of detailed work exploring the use of dual character statements in embodying-values cases. This work indicates that people think it is acceptable to use these statements in deeper cases that arise for concepts like art (Liao et al., 2020), law (Almeida et al., 2023; Almeida, forthcoming), scientist (Knobe et al., 2013), and even the concept human (B. Phillips, 2022).

Strikingly, one can also use dual character statements for embodying-values cases, legitimacy cases, and true-self cases. First, consider a legitimacy case. We might describe the robber baron using almost exactly the same words:

(18) There is a sense in which he is clearly the owner of this land, but ultimately, you would have to say that there is a deeper sense in which she is not the owner at all.

Or suppose we turn to a true-self case. We might say:

(19) There is a sense in which she is clearly happy, but ultimately, you would have to say that there is a deeper sense in which she is not happy at all.

The fact that we can use dual character statements in each of these cases provides at least some initial evidence for the claim that there is an important similarity between them.

Moreover, a closer look at the details of the results from existing studies gives us additional evidence that there is a real similarity here. In many of these studies, participants are not just asked whether they agree with the entire dual character statement as a whole. Instead, they are given each conjunct separately and asked to rate their agreement with each conjunct. For example, instead of being given statement (16) as a whole, they would be asked to separately rate their agreement with the two statements:

(20) a. There is a sense in which this is clearly art.
b. Ultimately, you would have to say that there is a deeper sense in which it is not art at all.

If participants agree with both of these statements, then they are agreeing with the dual character statement as a whole.

We can now look at the joint distribution of responses across these two statements. A number of studies have done this for embodying-values cases, and the results consistently indicate that the joint distribution shows the same distinctive pattern. To give one example, Figure 1 shows the joint distribution in a study (Phillips & Plunkett, 2023) that looked at responses to these two statements across different embodying-values cases (using the terms 'scientist,' 'friend,' 'artist,' etc.).

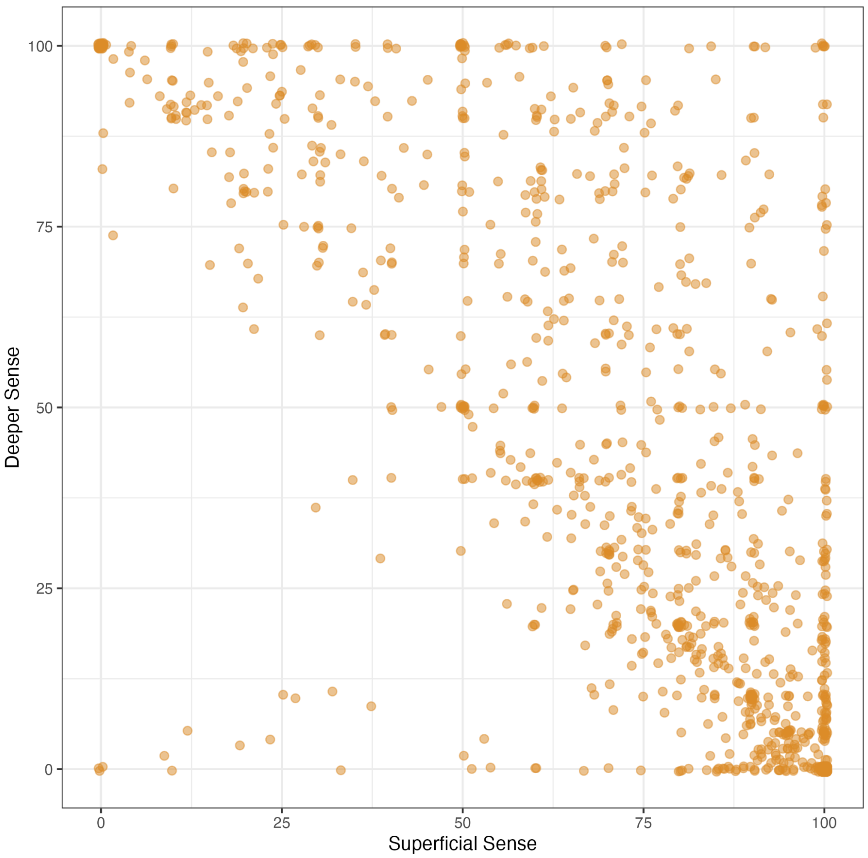

Figure 1. Ratings for the superficial sense statement and deeper sense statement in Phillips and Plunkett (2023). Figure generated using data that the authors made freely available in an online repository.

Looking at this figure, we see two things. First, there is support for the simple point made above. Of the participants who say that the term does not hold in a deeper sense (points in the upper part of the plot), many also agree that the term does hold in a more superficial sense (right side of the plot).

But, secondly, it's not as though all participants are in the upper-right quadrant; instead, we see a more complex pattern (negative correlation with high heteroskedasticity). It is not known why this pattern arises. In fact, as far as I know, there have been no attempts to offer a specific explanation of it.

Now consider legitimacy cases. In a recent study (Zhang et al., unpublished data), participants were given vignettes in which one group of people (e.g., the 'Gorps') steals something from another group of people (e.g., the 'Daxes'). Then participants were asked to rate their agreement with statements of the form:

[*Deeper*] Ultimately, the Daxes are the owners of the sea caves.
[*Superficial*] In a certain sense, the Gorps are the owners of the sea caves.

The joint distribution of these agreement ratings is shown in Figure 2.

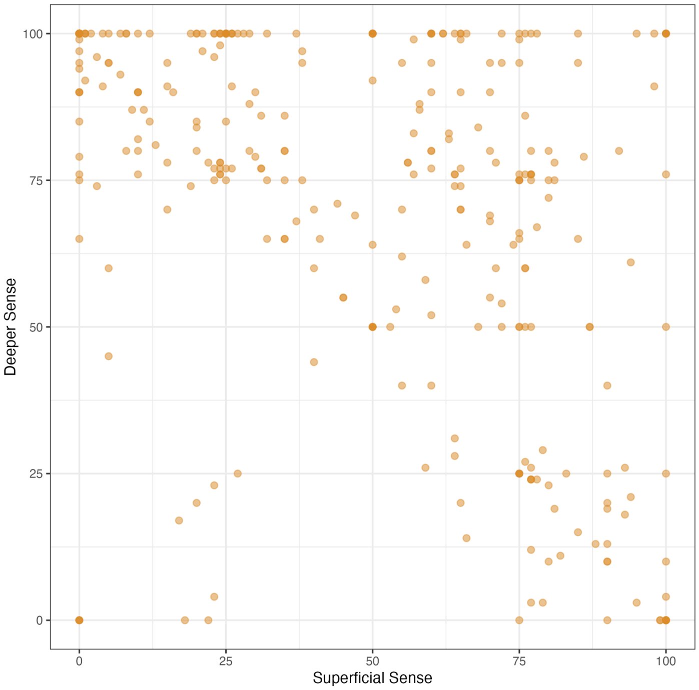

Figure 2. Ratings for the superficial sense statement and deeper sense statement in Zhang et al. (unpublished data).

In other words, the distribution observed for legitimacy cases appears to be quite similar to the distribution for embodying-values cases. I am not sure what to make of this finding. On one hand, it is not known what this distribution means, so it might seem inappropriate to draw any theoretical conclusions from the fact that it arises in both of these types of cases. On the other hand, it does seem striking that the distributions are so similar, and one might think that this fact provides further evidence for the claim that these different types of cases are similar.

3.5. *Normative significance*

Suppose someone reads this paper – the very paper you are reading right now – and says: 'This is not philosophy.' What I have been trying to suggest thus far is that such a person does not necessarily mean to deny that:

(21) There is clearly a sense in which this paper is philosophy

But that the person means to assert:

(22) In a deeper sense, this paper is not truly philosophy at all.

But now we can ask a further question: What exactly is a speaker trying to convey when she asserts (22) without denying (21)?

Clearly, a person who says this is not just trying to categorize the paper in a particular way. She is saying something more normative. She is trying to disparage the paper, to say that something has gone wrong in it. In previous work, this type of statement has been referred to as a *dual character diss* (Knobe, 2022).

This phenomenon appears to arise very generally for embodying-values cases. Consider what might happen if another person is clearly your friend in some superficial sense and you say: 'She is not really my friend.' Or suppose that something is clearly a work of art in some superficial sense and you say: 'This is not really a work of art.' In all of these cases, you are not just asserting that an object does not fall into a particular category. You are saying that something has gone wrong.

What is distinctive about the normative significance of these statements is that it arises from a conjunction of two claims, such that neither claim involves anything bad or wrong just in itself. The badness only arises from the whole conjunction. For example, there is nothing bad in itself about something not being philosophy. Lots of texts are not philosophy, and that doesn't mean that there is anything wrong with them. The thing that's bad is something that is both (a) clearly philosophy and (b) not philosophy in a deeper sense. When a text has both of those qualities, it is clear that something has gone wrong. Similar remarks apply in each of the other cases. There is nothing intrinsically wrong with not being your friend (most people are not your friend), but when a person is your friend in a superficial sense but is not your friend in a deeper sense, then it seems that something has gone wrong.

It might initially be thought that this phenomenon involves something quite specific to embodying-values cases, but that appears not to be true. Instead, this seems to be a very general feature of deeper-sense cases. First, consider legitimacy cases. Suppose we are talking about the robber baron who stole the indigenous people's land, and we say: 'In a deeper sense, he is not the owner of this land.' Here too, we are clearly trying to say that something has gone wrong, and in a very distinctive way. It's not as though there is anything wrong in itself with not being the owner of the land in a deeper sense. What is wrong is just being the owner in a more superficial sense but not being the owner in a deeper sense.

Finally, consider true-self cases. Suppose that there is a sense in which a person is clearly in love with someone, and we now say: 'In a deeper sense, she doesn't truly love him.' Here again, we see the same phenomenon. There is nothing wrong with not being in love with a particular person in a deeper sense, but there is something wrong with being in a state where there is a sense in which you are clearly in love but a deeper sense in which you are not.

This point seems to suggest one possible way of understanding the connection between the different types of deeper sense cases. Specifically, it might be suggested that these different cases are seen as similar because they involve something going wrong in a similar way. Take the case of a person who is king in a superficial sense but is not the legitimate king vs. the case of a person who is in love in a superficial sense but does not feel love deep down in her true self. Although these two cases are obviously different in numerous important respects, one might think that they involve things going wrong in the same way. To give a first stab at what is similar across all types ol deeper sense cases, one might say something like: 'This object is failing to be the thing that it pretends to be.' Of course, this characterization is far too nebulous to count as a real hypothesis, and I have not been able to find a way to spell it out more clearly. But this does seem to be a possible avenue for further research.

3.6. *No fixed list*

Thus far, we have been considering certain specific expressions that can be understood both in a superficial sense and in a deeper sense: 'scientist,' 'king,' 'happiness,' and so forth. This framing may create the impression that there is a fixed list of expressions that have this feature and that all other expressions cannot be understood in these two different senses. But, importantly, existing research shows that this is not the case (Baumgartner, forthcoming; Phillips & Plunkett, 2023). Rather, it seems that if a speaker thinks that things have gone wrong in a particular way, that speaker can just spontaneously begin using a term in two different senses, even if it had never been used in that way before.

To illustrate, consider a new sort of legitimacy case. At the moment, I don't think most people would think that the term 'United States Transportation Secretary' has two distinct senses. But suppose that something unexpected happens: A power-hungry railroad magnet launches a violent coup. She has the duly-appointed transportation secretary kidnapped and then begins occupying the role of transportation secretary herself. In such a case, defenders of the rule of law might suddenly begin saying things like: 'Ultimately, she is not the true transportation secretary.' To the extent that we do not see this expression as having two senses right now, this might just be a matter of the fact that we have never encountered a case in which things have gone wrong in this way.

The same point applies to embodying-values cases. At the moment, I would not say that the phrase 'experimental philosopher' has two different senses, and I cannot think of any occasion on which someone said something like: 'Ultimately, she is not a true experimental philosopher.' But this is not because there is anything built into the very semantics of that expression that makes this impossible. If we ever encounter a case in which we do think that someone is failing to embody the relevant values, we would then say precisely that. (Imagine that corporations begin hiring people to use the tools of experimental philosophy in projects aimed at marketing consumer products.)

And the same can be said of true-self cases. At the moment, I would not see the word 'boredom' as having two senses, but that is not because there is something built into the semantics of this word that makes it impossible to understand a notion of 'true boredom.' We can imagine things going wrong in such a way that we would say that certain people feel bored on a superficial level but are not truly bored deep down in their true selves, and if that did happen, we would describe such people by saying: 'Ultimately, this is not true boredom.'

Cases like these seem to point to something about what is involved in certain terms having two distinct senses. At least initially, one might be tempted to understand this in terms of people literally associating these terms with two distinct representations (a representation of the superficial sense, a representation of the deeper sense). But cases like these make that picture look implausible. It is not very plausible that people have two different representations for the meaning of the phrase 'Secretary of Transportation' and it just happens that most people think that these two representations have picked out the same person in all cases that have arisen thus far. Instead, these cases seem to point to a picture that looks at least broadly like this: People have a representation of the meaning of this term. Then that representation has the property that if things go wrong in a certain way, people will be inclined to think that the term applies to an object in one sense but not in another. Further theoretical research will be needed to spell out this picture in more detail.

3.7. *Summing up*

This section has explored six different respects in which embodying values cases, legitimacy cases and true self cases appear to be strikingly similar.

1. People use the same linguistic expressions to pick out the deeper sense ('true,' 'ultimately,' 'in a deeper sense').

2. When people do not use any such expression, it is not clear whether they intend a term to be understood in the superficial sense or the deeper sense.

3. Introducing the deeper sense does not involve rejecting the superficial sense; the two can coexist within a single conversational context.

4. The superficial sense and the deeper sense can be explicitly used together in a 'dual character statement.'

5. When people say that a term holds in a superficial sense but not in a deeper sense, they seem to be saying that something has gone wrong.

6. There is no fixed list of terms that can be understood in both senses. It is always possible to begin using a new term in this way if one encounters a new way in which things can go wrong.

In light of all of this, it seems that we have at least some reason to reject theories on which these are just three completely separate types of cases, and that we have at least some reason to prefer theories according to which they are best understood as different manifestations of the same underlying phenomenon.

4. Conclusion and remaining questions

I have been arguing that people have a single unified notion of a term not applying to an object 'in a deeper sense.' Embodying-values cases, legitimacy cases, and true-self cases can then all be understood in terms of this one notion.

If this claim is correct, the notion of a 'deeper sense' is far more pervasive than one might initially have thought. One might have thought that this notion only applied in those specific cases where people are wondering whether an object embodies the values that characterize a category and that in all other cases it simply is not relevant. But the evidence seems to suggest that this is not correct. Instead, this notion appears to apply far more broadly, across a very wide range of cases.

Perhaps more importantly, if this view is correct, then the notion of an object failing to embody the characteristic values of a category is not truly essential to the notion of a term failing to apply in a deeper sense. It just happens that researchers began exploring the latter notion by looking at cases that involved the former. But the notion of a term failing to apply in a deeper sense seems to be something far more general. If we want to understand it, we will have to explore it at this more general level.

4.1. *Revisiting the differences*

One obvious question raised by this view is how to understand the apparent differences between different types of cases. It certainly seems that there is something very different between saying that someone is not the king in a deeper sense and saying that someone is not happy in a deeper sense. How are we to make sense of these differences?

I do not know the answer, and this is certainly an area in which further research is needed. However, one tempting hypothesis would be that the differences, though real, are smaller than they might at first appear. In particular, it might be that they might be best understood as differences in degree rather than differences in kind.

To illustrate this hypothesis, consider again the expression 'true king.' I suggested above that judgments about whether someone counts as a true king will be determined primarily by whether he acquired his status through a legitimate process (e.g., being the son of the previous king). But now we can distinguish two different ways in which this pattern of judgments might be explained. One possibility would be that when people learn the meaning of the phrase 'true king,' what they learn is that it means something like *legitimate king*. Alternatively, a second possibility would be that when people learn the meaning of this phrase, what they learn is just that it means something like *king in a deeper sense*. Then it might be a further fact – not built into the very meaning of the expression – that when people are trying to determine whether someone is king in a deeper sense, they mostly tend to focus on legitimacy.

If we think that the difference between different types of deeper-sense cases is indeed best understood in this way, then a natural further hypothesis would be that this difference is just *a matter of degree*. People's judgments about whether someone is a 'true king' might be determined primarily by questions about legitimacy, but that would not mean that no other considerations play any role. Instead, these judgments might also be impacted at least to some degree by questions about embodying the characteristic values of being a king, such as wisdom or courage (embodying values or by questions about feelings called to be a king by something deep within himself (true self). In other words, it might be that we have a notion of being king in a deeper sense is *mostly* a matter of legitimacy but that is also partly a matter of embodying values and true self.

If this hypothesis turns out to be correct, we might need to revisit certain claims from the existing empirical literature. Existing studies suggest that judgments about whether someone is a true scientist are based on whether the person embodies the values that are characteristic of being a scientist (Knobe et al., 2013), whereas judgments about whether someone is experiencing true love are based on whether the love is rooted in the true self (Prinzing et al., 2023). But perhaps this is just a matter of degree. It might be that judgments about whether someone is a true scientist are also impacted at least in some small way by information about the true self (see Del Pinal & Reuter, 2017) and that judgments about whether something is experiencing true love are impacted at least in some small way by whether the person embodies the values that are characteristic of love.

4.2. *Explaining the similarities*

Finally, and most importantly, we face a question about what exactly it is that unites the different types of deeper-sense cases. What is it about embodying-values cases, legitimacy cases and true-self cases that makes people see all of them as similar and understand all of them in terms of a distinction between a superficial sense and a deeper sense? Here again, I do not know the answer, but I want to close by considering an approach we might pursue in trying to figure this out.

As I noted earlier, it seems plausible that all types of deeper-sense cases involve something going wrong in a similar way. In particular, in all of these cases, we might be inclined to say something of the form: 'This object is failing to be the thing it pretends to be.' Or, putting it slightly differently, we might be inclined in all cases to say that the object is *fake*. Thus, one possible avenue for future research would be to explore the notion of fakeness and how that notion might relate to each type of deeper-sense case.

As an initial step in developing this approach, let's consider a more straightforward case. Consider a children's toy that has been designed to look like a gun. We might think that there is some way in which it is pretending to be a gun but is not actually a gun, and we might therefore refer to it as a 'fake gun.' (For existing work on phrases like this one, see, e.g., Coulson & Gilles, 1999; Pavlick & Callison-Burch, 2016.) Clearly, there are some important differences between our deeper-sense cases and this fake gun case, but the suggestion we will be exploring is that there are also some important similarities.

Note to begin with that we might describe the fake gun case using similar linguistic expressions to the ones we would use in deeper-sense cases. For example, there does seem to be a sense in which this object can be described as a gun ('Hand me that gun'), and yet, at the same time, there is also a sense in which it can be said not to be a gun at all ('This is not a gun'). Moreover, we can get at that latter sense by using the word 'real' ('This is not a real gun'). Work in linguistics has therefore already proposed that embodying-values cases should be seen as similar to the case of a fake gun when it comes to their more abstract logical properties (Del Pinal, 2018).

The key suggestion now is that what all deeper-sense cases have in common is that people see all of them as cases of fakeness. So although each type of case might be seen as involving something very different, people see all of them in terms of an object being fake. The scientist who fails to embody what science is really all about is seen as a 'fake scientist'; the king who is not legitimate is seen as a 'fake king'; and happiness that is not rooted in the true self is seen as 'fake happiness.'

Obviously, these brief remarks are still very far from constituting a testable hypothesis, but perhaps they are nonetheless sufficient to constitute a strategy for further research. At a theoretical level, the key task is to articulate an understanding of what it would mean for people to see all of these cases as involving an object being fake, i.e., failing to be what it pretends to be. Then, at an empirical level, the task would be to develop and apply a method for determining whether people do indeed see all of these cases in that way.

References

Almeida, G., Struchiner, N., & Hannikainen, I. R. (2023). Rule is a dual character concept. *Cognition*, 230, 105259.

Almeida, G. (forthcoming). A dual character theory of law. *Journal of Legal Philosophy*.

Bailey, A. H., Knobe, J., & Newman, G. E. (2021). Value-based essentialism: Essentialist beliefs about social groups with shared values. *Journal of Experimental Psychology: General,* 150(10), 1994.

Barker, C. (2002). The dynamics of vagueness. *Linguistics and Philosophy*, 1-36.

Baumgartner, L. (forthcoming). The pragmatic view on dual character concepts and expressions. *Mind & Language*.

Coulson, S. & Gilles F. (1999). Fake guns and stone lions: Conceptual blending and privative adjectives. In B. Fox, D. Jurafsky & L. Michaels (eds.), *Cognition and function in language*. Palo Alto, CA: CSLI.

Del Pinal, G. (2018). Meaning, modulation, and context: a multidimensional semantics for truth-conditional pragmatics. *Linguistics and Philosophy,* 41, 165-207.

Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, 41, 477-501.

Demaree-Cotton, J., & Sommers, R. (2022). Autonomy and the folk concept of valid consent. *Cognition*, 224, 105065.

Guo, C., Dweck, C. S., & Markman, E. M. (2021). Gender categories as dual‐character concepts?. *Cognitive Science*, 45, e12954.

Khoo, J. (2020). Quasi indexicals. *Philosophy and Phenomenological Research*, 100(1), 26-53.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127, 242-257.

Knobe, J. (2022). Personal identity and dual character concepts. *Experimental Philosophy of Identity and the Self*, 49.

Leslie, S. J. (2015). 'Hillary Clinton is the Only Man in the Obama Administration': Dual Character Concepts, Generics, and Gender. *Analytic Philosophy*, 56.

Liao, S. Y., Meskin, A., & Knobe, J. (2020). Dual character art concepts. *Pacific Philosophical Quarterly*, 101, 102-128.

Malone, E. (2023). Country music and the problem of authenticity. *British Journal of Aesthetics*, 63, 75-90.

Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, 17, e12823.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, 39(1), 96-125.

Pavlick, E. & Callison-Burch, C. (2016). So-called non-subsective adjectives. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics* (pp. 114-119).

Phillips, J., & Plunkett, D. (2023). Are there really any dual‐character concepts? *Philosophical Perspectives*, 37, 340-369.

Phillips, B. (2022). "They're not true humans": Beliefs about moral character drive denials of humanity. *Cognitive Science*, 46(2), e13089.

Plunkett, D., & Sundell, T. (2013). Disagreement and the semantics of normative and evaluative terms. *Philosophers' Imprint* 13, 1-37.

Prinzing, M., Earp, B. D., & Knobe, J. (2023). Why do evaluative judgments affect emotion attributions? The roles of judgments about fittingness and the true self. *Cognition*, 239, 105579.

Prinzing, M. M. & Fredrickson, B. L. (2023). No peace for the wicked? Immorality is thought to disrupt intrapersonal harmony, impeding positive psychological states and happiness. *Cognitive Science*, 47, e13371.

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, 14(1), e12557.

Zhang, F., Chang, J., Prinzing, M & Knobe, J. (unpublished data). Intuitions about ownership in cases of illegitimate transfer. Yale University.