Synthese (2023) 202:16
https://doi.org/10.1007/s11229-023-04221-3

ORIGINAL RESEARCH

Do you believe in Deep Down? On two conceptions of valuing

Marcel Jahn[1] · **Lukas Beck**[2]

Received: 4 November 2022 / Accepted: 4 June 2023 / Published online: 5 July 2023

Abstract

In this paper, we explicate an underappreciated distinction between two conceptions of valuing. According to the first conception, which we call the surface-account, valuing something is exclusively a matter of having certain behavioral, cognitive, and emotional dispositions. In contrast, the second conception, which we call the layer-account, posits that valuing is constituted by the presence of certain representational mental states underlying those dispositions. In the first part of the paper, we introduce the distinction in proper detail and show that the accounts have different implications regarding the valuings of agents. In the second part, we situate the accounts within the extant philosophical literature. First, we relate them to the recent debate between so-called dispositionalists and representationalists about the nature of beliefs and point out that this debate can help anticipate some of the main dialectical fault lines to be expected between surface- and layer-theorists. Second, we examine the contemporary meta-ethical debate on conceptualizing valuing, indicate that scholars have largely ignored the distinction introduced here, and outline that this oversight has substantial theoretical costs: as we show, key arguments within the meta-ethical debate require thorough re-evaluation in light of the proposed distinction. The third part of the paper illustrates the theoretical leverage of the distinction for practical research by exploring its implications for behavioral welfare economics.

Keywords Valuing · Meta-ethics · Behavioral welfare economics · Representationalism · Dispositionalism

✉ Marcel Jahn
marcel.jahn@hu-berlin.de

Lukas Beck
beck@mcc-berlin.net

[1] Institut für Philosophie, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

[2] Mercator Research Institute on Global Commons and Climate Change (MCC), Torgauer Straße 12-15, 10829 Berlin, Germany

1 Introduction

BoJack: *"Well, do you think I'm a good person… deep down?"*
Diane: "That's the thing. I don't think I believe in deep down. I kind of think all you are is just the things that you do."
BoJack Horseman, Episode 12, Season 1, 21:53

Many scholars agree that valuing something entails having various dispositions (see Schwitzgebel, 2002; Kolodny, 2003; Scheffler, 2004; Wallace, 2013). For instance, valuing a romantic partnership plausibly involves dispositions such as being willing to compromise your career or to make serious efforts to resolve tensions. In addition, you will generally be disposed to respond with appropriate emotional reactions to the ups and downs in your partner's life. Moreover, when you deliberate what to do, you will be disposed to give special weight to your partner's needs.

Given the centrality of dispositions to the phenomenon of valuing, how should we judge cases in which some of these characteristic dispositions are absent? Should we say that the agents in these cases still have the valuing in question? For example, let us imagine Barak, who is currently very depressed. Some time ago, he had all the characteristic dispositions associated with valuing one's partnership. Since his depression, however, things have changed. He does not commit effort to resolve tensions with his partner Abby, shows little emotional response towards the fortunes and misfortunes in Abby's life, and lacks the energy to give his partner's needs any special significance in his practical deliberations. Yet, even though Barak lacks these characteristic dispositions for valuing a romantic partnership, he would still assert that he truly values his partnership.

In this paper, we argue that cases like this one reveal the need to distinguish between and work out two accounts of valuing. According to the first, which we refer to as the surface-account, valuing something is exclusively a matter of having certain behavioral, emotional, and cognitive dispositions. In contrast, the second account, which we call the layer-account, posits that valuing is constituted by the presence of certain representational mental states underlying those dispositions. The main difference between these accounts is that the first account *identifies* valuing with having certain dispositions, while the second account treats those dispositions as *mere effects* of valuing.

The distinction between these two accounts has received almost no explicit recognition in the philosophical literature. However, we believe that said distinction marks a crucial watershed in how we should think about valuing and that failure to acknowledge it comes at considerable theoretical costs. As we will show, ignoring this distinction has undesirable consequences both within philosophy as well as beyond philosophy. Our goals are, therefore, to draw attention to the distinction, point out the multifaceted theoretical benefits that it offers, and emphasize the need to decide whether one is committed to the surface- or the layer-account.

The paper consists of three main parts. In the first part, we will lay out the accounts in sufficient detail and explore their diverging implications (Sect. 2). In the second part, we will embed the distinction within the current philosophical literature. On the one hand, we will relate it to a debate within the philosophy of cognitive science

where a similar distinction has recently gained prominence. Specifically, *representationalists* (see, e.g., Quilty-Dunn and Mandelbaum, 2018) and *dispositionalists* (see, e.g., Schwitzgebel, 2001, 2002, 2010, 2013, 2021) argue about whether beliefs should be conceptualized as representational states or as dispositions. We will illustrate that this controversy on the nature of beliefs already reveals some of the main dialectical fault lines to be expected in a debate about the valuing-case and point out some of the peculiarities that make it particularly difficult to navigate (Sect. 3). On the other hand, we will relate the proposed distinction to the contemporary meta-ethical debate on conceptualizing valuing, suggesting that it has received almost no attention here and that this is a consequential oversight. In particular, we will point out that scholars who analyze the concept of valuing run the danger of talking past each other by neglecting the distinction – a danger that has indeed materialized in practice. This calls for a re-assessment of central arguments within the debate and, we argue, a vigorous engagement with the question of whether to opt for the surface- or the layer-account (Sect. 4). In the third and final part, we will examine the implications of recognizing the proposed distinction for practical research, arguing that it grants substantial theoretical leverage in fields of social scientific research that implicitly rely on some understanding of valuing. To make this case, we will examine behavioral welfare economics specifically and suggest that without explicit recognition of the distinction, key meta-normative assumptions within this field remain unrecognized: as we will illustrate, most work in this area is premised on a layer-conception of valuing. Yet, we will suggest that there is an underexplored – yet quite promising – way of conducting behavioral welfare economics based on a surface-conception of valuing (Sect. 5). All in all, then, putting the distinction between the surface- and the layer-account center stage can lead to major progress, at least in the current meta-ethical discourse and areas of practical research such as behavioral welfare economics.

2 Two accounts of valuing

Let us begin and introduce the two accounts. To this end, we will first motivate them by presenting the case of Barak in more detail, then define the accounts, and finally return to the case of Barak to illustrate their differences.

2.1 The case of Barak

Imagine Barak, who is in a partnership with Abby. Barak appears to be a very decent partner who truly values his relationship. As such, he commits substantive effort to resolving upcoming tensions, he is empathetic to the ups and downs in his partner Abby's life, and he takes the needs of Abby very seriously in his decision-making. As the years pass, however, more and more difficulties emerge in Barak's life: he increasingly gets dragged down by the challenges he faces at work, he loses sight of dearest friends, and, as he ages, some unpleasant health issues arise. Initially, all of these factors do not affect how Barak behaves towards Abby.

However, after some time, we start to notice that he begins to forget about Abby's needs, is emotionally less affected by the ups and downs in Abby's life, and some-

times cannot muster the strength necessary for working through tensions. In other words, we can observe that some of Barak's dispositions associated with valuing his partnership start to weaken. These tendencies only get stronger over time. At some point, Barak and Abby have a series of conversations about their relationship given how much Barak seems to have changed. They are confident to work through this hard phase and Barak asserts that, even though times are hard for him, he still values their partnership. As this conversation indicates, at this point in time, Barak still has (at least to a certain extent) a disposition to work through tensions. Yet, the reason why the two have the conversation in the first place is that Barak lacks, or only has to a small degree, several of the other dispositions characteristic of valuing one's partnership.

Nevertheless, things only get worse after their conversation. Eventually he becomes seriously depressed and lacks almost all of the dispositions characteristic of valuing a partnership. Yet, he still holds that he values his partnership and would, if asked, sincerely assert this.[1]

How should we evaluate this case? It is intuitively very plausible to suppose that Barak valued his relationship in the beginning, before any difficulties entered his life. And even at the time when he had the conversation with Abby, it may still be intuitive to say that Barak values the partnership. However, it seems particularly unclear what we should say when Barak is in the midst of his depression. At this point, one may have one of the following two intuitive reactions, both of which seem quite understandable. On the one hand, one might hold that Barak no longer values his partnership because he almost completely lacks the characteristic dispositions associated with valuing his partnership. In contrast, one might be inclined to judge that he still

[1] Of course, one could say that in our example, Barak's dispositions are merely masked (see Bird, 1998), i.e., that Barak still has the relevant dispositions even though they do not manifest. To illustrate, one could, for instance, imagine a case in which Barak's depression disappears after taking a drug and all the dispositions characteristic of valuing come back. If this were so, it may appear plausible to construe Barak's case as one of masking. As a reply, let us highlight that our exposition of the surface-account is, as we will also point out below, committed to the so-called simple conditional analysis of dispositions. This account deals with the alleged masking case by distinguishing two different dispositions based on their stimulus conditions, e.g., the disposition to give special weight to your partner's needs versus the disposition to give special weight to your partner's needs *while not being depressed* (see Choi, 2006). We can then say that Barak lacks the former disposition, which we consider essential for valuing. At this point, however, one could argue that the dispositions that are relevant for valuing one's partnership are distinct from the dispositions one can have when depressed, e.g., the relevant disposition is the disposition to give special weight to your partner's needs *while not being depressed.* Hence, one could claim that we misspecify the dispositions relevant for valuing one's partnership. In response, we admit that which dispositions one takes to be characteristic of valuing something is itself a contentious issue even among those who accept a surface-account of valuing. Yet, these contentions do not threaten the core distinction we are concerned with in this paper. Lacking the space to outline all relevant details of more complex cases, let us note that we use the example of Barak's depression primarily as an intuition pump aimed at highlighting the existence of vexed cases in which the two accounts will lead to different verdicts. Suppose one identifies a different set of characteristic dispositions for valuing one's partnership than we do. In that case, one is free to run a more complex example, e.g., one in which Barak is not displaying the characteristic dispositions for other reasons than being depressed, while still asserting that he truly values his partnership. For instance, Barak may simply be so knocked-down and broken that he only has a few dispositions relevant for valuing a partnership left. Therefore, motivating the distinction does not require positing that Barak has a clinical depression. We thank an anonymous reviewer for pressing us on this issue.

values his partnership; after all, one might argue, he still represents his partnership as a valuable aspect of his life, which is evidenced by his sincere assertion.

At first glance, it is unclear which of these reactions should be favored. But in any case, they reveal two very different understandings of what it is to value something. Crucially, these two understandings are not unique to personal relationships, but represent two general conceptions of valuing, applicable to any object of valuing such as parts of nature, health, sunsets, or walks in the park.[2]

2.2 The surface-account

What we call the *surface-account* emerges from taking seriously the idea that having certain dispositions is the essential feature of valuing. This account defines valuing exclusively as a matter of having dispositions that are characteristic of valuing the item in question.

> *Surface-Account.* What it is for an agent A to value φ is for A to have the characteristic dispositions for valuing φ to a sufficient degree.

Let us make three clarificatory remarks about this account. First, what do we mean by dispositions? Roughly speaking, a disposition is supposed to be a feature of an object that is not actually observable but manifests only if certain stimulus conditions are present. For instance, that sugar is (water-)soluble is a property of sugar that shows up only if sugar is put into water and thereby dissolves. Agents can have dispositions as well. Tom may have a disposition to pick up his children after school; this disposition manifests itself in his driving to school, putting his children in the back seat of his car, and driving them home, assuming it is 2pm on a weekday (the time school ends).

More generally, we here characterize dispositions by reference to the following conditional statement that Eric Schwitzgebel (2002, p. 250), who we think would be generally sympathetic to the surface-account, formulates as follows: "[if] condition C holds, then object O [or agent A] will (or is likely to) enter (or remain in) state S."[3] Here, condition C denotes the stimulus condition and the fact that the object O (or the agent A) enters state S denotes the disposition's manifestation.

Now, there is a lively debate in metaphysics about how to exactly analyze dispositions in terms of conditionals, and in particular whether the so-called *simple conditional analysis* just mentioned is ultimately successful. Moreover, philosophers debate the question of how to spell out precisely the truth conditions of the conditional statements involved in these conditional analyses (see Prior, 1985; Choi and Fara, 2021).[4] Yet, for our purposes, it does not really matter how these specifics are sorted out. The reason is that we take the simple conditional analysis of disposi-

[2] We thank an anonymous reviewer who encouraged us to highlight the general character of the two conceptions.

[3] See also Manley and Wasserman (2008, 60).

[4] There are also philosophers who defend the idea that one can analyze dispositions without linking them to conditionals (see Vetter, 2015).

tions to offer the intended reading of 'dispositions' in the formulation of the surface-account above. As we will see below, this conception allows us to best highlight the differences between the surface- and the layer-account.[5] Moreover, as outlined above, authors like Schwitzgebel (2002), who have defended dispositional accounts of belief, also stipulate and work with the simple conditional analysis. However, suppose one is committed to a different conception of dispositions and, therefore, unhappy with our use of the term here. In that case, one is free to drop the term 'dispositions' and speak of patterns in our acting, thinking, and feeling instead.[6]

This brings us to the second clarificatory remark. We hold that in the case of valuing, there are usually three types of dispositions in play: behavioral, emotional, and deliberative dispositions. To remind you of the case that we started with, valuing a romantic partnership characteristically involves, say, a willingness to compromise one's career (behavioral disposition), being emotionally vulnerable to the ups and down in one partner's life (emotional disposition), and placing special significance to one partner's needs in deliberating what to do (deliberative disposition).

Finally, let us clarify what we mean by having the characteristic dispositions for valuing φ *to a sufficient degree*. The qualification "to a sufficient degree" is important because intuitively, valuing something does not entail that one possesses all characteristic dispositions to a perfect degree. An agent might, say, lack a certain disposition that is characteristically associated with valuing one's partnership, which need not mean that she does not value her partnership. However, once the agent lacks a sufficiently large subset of the characteristic dispositions, one might argue that she does no longer value her partnership. Or, one could imagine that the agent lacks an entire class of dispositions (e.g., she has the relevant behavioral and deliberative dispositions, but none of the emotional dispositions). In this case, too, one could conclude that she does not value her partnership.

Agents may fail to exhibit a sufficient degree of the characteristic dispositions not only due to an insufficient number of dispositions, but also due to the fact that dispositions are weakly pronounced. For example, it could be that, for many of the dispositions, there is a relatively small likelihood that the respective behavior occurs when the relevant stimulus condition obtains. To illustrate, imagine an agent who is only weakly disposed to place special weight on their partner's needs in practical deliberation, i.e., they rarely take their partner's needs into account. Now, minor or moderate qualitative failings may be compatible with valuing, but not if they are sufficiently severe. All in all, according to the surface-account, valuing does not necessitate the presence of all characteristic dispositions to a perfect extent, but once an agent possesses them only to an insufficient degree the agent fails to have the respective valuing.

[5] As will be explained below, it underlines that the surface-account expresses a decidedly practically viewpoint aimed at "how people will act when it matters," instead of providing conceptual space for (spurious) excuses (Schwitzgebel, 2021, p. 363).

[6] In other words, one may speak of input-output relations between stimuli and the agent's acting, thinking, and feeling.

2.3 The layer-account

As we just saw, the surface-account holds that what it is to value something is to have certain characteristic dispositions (to a sufficient degree). In contrast, the *layer-account* states that valuing is constituted by the presence of a certain representational mental state.

> *Layer-Account.* What it is for an agent A to value φ is for A to have a particular (cognitive or conative) mental state ψ that represents φ as valuable.

At first glance, it looks as if the layer-account does not acknowledge that dispositions play a role in valuing. Given the centrality of dispositions to the phenomenon of valuing, it would be a highly dubious and bold implication of the layer-account if this was indeed the case. Crucially, however, proponents of the layer-account need not deny that dispositions are an important aspect of valuing. They are simply saying that valuing – at its core – is not a matter of the presence of certain dispositions, but that valuing is essentially a matter of the presence of a particular representational mental state.[7] This is consistent with the presence of certain dispositions. A natural suggestion would be that, at least in paradigmatic cases of valuing, the relevant representational mental state causally contributes to the agent's having the characteristic dispositions.[8]

Thus, a proponent of the layer-account will be dissatisfied with the surface-account primarily because it stops its analysis of valuing at a too superficial level. She will argue that valuing φ is at bottom about the presence of a certain representational mental state and that the characteristic dispositions associated with valuing φ are merely effects of valuing φ. Consequently, the presence of these dispositions can serve as evidence for whether the agent has a certain value, but – and this is the crucial point – they are not constitutive of valuing.

[7] Note that one could, of course, also analyze mental representations in terms of their dispositional or functional characteristics. Functionalists about mental states even maintain that something is not a mental state in virtue of any of its categorical features but instead in virtue of its function or dispositional characteristics. However, functionalists about *representational* mental states usually specify their role in terms of "functions within the mind that often bear only remote connections to stimuli, behavior, and phenomenology" (Quilty-Dunn & Mandelbaum, 2018, p. 2354). Thus, these functions are clearly different from the agent-level dispositions that the surface-account identifies with valuing. In this regard, one should not misread the surface-account as a rather coarse-grained functionalist account of the r*epresentational* mental state of valuing. Understood like this, it would offer a rather implausible take on mental representations as it will be hard to identify any proper sub-system of the agent that could be viewed as the causal base of *all* the relevant dispositions. Saying that the whole agential system (or a very large part of it) constitutes a mental representation makes for an implausible candidate for mental representations. Hence, a functional analysis of mental representations does not threaten the distinction between the surface- and the layer-account.

[8] One could also hold the view that both having the representational mental state and the dispositions are necessary for valuing. Yet, such an account could not deliver the judgment that Barak values his relationship even when he lacks the relevant dispositions; thus, it could not cater to the intuitions by proponents of the layer-account. Moreover, this rather complex account would also lose out on the practical advantages of the surface-account (see Sect. 3.3 for more details). We therefore believe that said account does not provide any edge over either the surface- or the layer-account, but rather represents an overall more demanding account of valuing. Hence, adopting this account does not seem particularly motivated.

To further illustrate the layer-account, it is helpful to clarify what it does *not* take a position on. First, the layer-account remains agnostic on whether the representational mental state relevant for valuing is cognitive (i.e., a belief-like state that represents states of the world) or conative (i.e., a desire-like mental state that represents certain goals).[9] Hence, the layer-account does not provide a comprehensive, fully specified view of valuing. Instead, it subsumes a family of such comprehensive views, all of which assume that valuing is primarily a matter of having a particular representational mental state.

Now, depending on how one develops this detail of the layer-account, we obtain different variants of it. If we flesh out the layer-account with recourse to cognitive mental states, "representing φ as valuable" means representing φ as something that has certain valuable features. In this vein, Smith (1992, p. 344) suggests that "valuing is believing valuable," where believing valuable is "believing that we have normative reason."

If, on the other hand, we were to specify the layer-account by reference to conative mental states, then "representing φ as valuable" would mean representing φ as particularly worthy of being implemented or maintained. There are some philosophers who have expressed sympathy for an analysis of valuing along these lines. Harman (2000, p. 135), for example, states that "valuing is a particular kind of desiring," although he adds that he is "unable to say more about what kind." Even more ambitious is Lewis (2000), who offers an analysis of valuing in terms of second-order desires, claiming that A values φ iff A desires to desire φ.[10]

In addition to not committing itself to whether the representational state in question is cognitive or conative, the layer-account also does not take a particularly firm stance on what representational mental states ultimately are. Broadly speaking, representational mental states are often understood as physical structures (usually assumed to reside in the brain) that represent their content and on which computational operations can be performed (see Pitt, 2022). Alternatively, under so-called role-functionalist accounts, mental representations can be identified with the second-order property of occupying a particular functional role (see McLaughlin, 2007). The layer-account, however, is not tied to any precise view of what makes it the case that

[9] For an example of the common cognitive-conative distinction of different representational mental states see, for instance, Schulz (2018, chs. 5–6).

[10] There are accounts of desire according to which having a desire simply means to have a single disposition to act in a particular way. According to this view, having a disposition to act is what it means to have a desire, and there are no other essential features to desiring (e.g., feeling or thinking in certain ways) (cf. Schroeder, 2020). Hence, these accounts of desiring are a form of what we would call thin-dispositionalism that characterizes an attitude (e.g., a desire) in virtue of a single (behavioral) disposition. This contrasts with the thick-dispositionalism of our presentation of the surface-account that characterizes valuing in terms of a whole zoo of dispositions. Moreover, thin-dispositional accounts of desire could still identify the causal base of the relevant disposition with the realizer of a (representational) mental state, which would turn them into a version of the layer-account because the surface-account is not intended to specify any mental representations. Yet, even if thin-dispositional accounts of desires are not meant to identify genuine mental representations, valuing, under the surface-account, is most plausibly not exhausted by a single disposition. Instead, it is more plausibly constituted by a large set of agent-level dispositions. Hence, under this latter construal, a thin-dispositionalist analysis of desires would not provide us with a plausible candidate of valuing on either the surface- or the layer-account (see also Heathwood (2019) on merely "behavioral desires").

something constitutes a mental representation. Moreover, it is also not tied to any assumptions about what makes it the case that a representational mental state has the content that it does (see Ramsey, 2016). The layer-account is merely bound to the claim that there are mental representations and that having a particular subset of them is constitutive of valuing (see also Spurrett, 2021).[11]

2.4 Revisiting the case of Barak

Having presented the surface- and the layer-account, let us return to the case of Barak. We will illustrate that depending on which stage of Barak's development we focus on, the accounts can lead to different judgments about Barak's valuings, and in cases where they do reach the same verdict, they do so for different reasons.

First consider the period when Barak and Abby's relationship was still intact and Barak had all the dispositions normally associated with valuing a partnership. For this period, both accounts are in agreement that he valued his partnership with Abby. However, they would reach this verdict for different reasons. According to the surface-account, Barak has the valuing simply because he has all the relevant dispositions. A proponent of the layer-account, on the other hand, would not directly infer that Barak has the valuing from the fact that he has these dispositions; rather, this fact merely functions as evidence that he has a particular mental representation that is constitutive of valuing his partnership. (Recall that this evidential relation obtains due to the assumed causal relation between the representational state and the dispositions.)[12]

The accounts can lead to opposing judgments, however, when we consider the time during which Barak is severely depressed. Since Barak lacks almost all of the dispositions characteristic of valuing one's partnership, the surface-account's verdict seems clear: he does not have the valuing in question.[13] In contrast, a proponent of the layer-account could argue for the opposite verdict. Barak values his partnership because – despite everything – he sincerely asserts that he still values the partnership. An advocate of the layer-account could treat this assertion as sufficient evidence for the fact that he still possesses the representational mental state constitutive of valu-

[11] In the context of the layer-account, we presuppose a realist understanding of representational mental states as opposed to an interpretivist view (see, e.g., Dennett, 1987).

[12] As indicated above, the two accounts are general accounts of valuing and thus not only applicable to valuing personal relationships. Let us briefly mention another example: valuing (parts of) nature. According to the surface-account, it would be constitutive of this valuing to have a sufficient degree of dispositions such as the following: being disposed to engage with nature in certain ways, to facilitate policies to protect nature, to feel sadness when one sees parts of nature being destroyed, and to take "nature's interest" into account when deliberating how to live one's life. According to the layer-account, it would be constitutive of this valuing to have some representational state, like a certain second-order desire directed at (parts of) nature.

[13] As mentioned in footnote 1, another way to challenge this verdict is to argue that we misspecify the dispositions relevant for valuing one's partnership. Again, we believe that how to correctly characterize the relevant dispositions is itself a contested issue. Still, for the sake of adequately illustrating the distinction we are concerned with here, and to avoid getting sidetracked by the independent issue of how to specify the dispositions relevant for valuing one's partnership, we invite the reader to grant us, for the moment, that we have specified them appropriately.

ing. That such sincere assertions do indeed provide sufficient evidence for a certain representational state is a position that many find quite appealing. For instance, numerous philosophers working in epistemology accept the view that the agent's sincere assertion "I believe that p" is sufficient evidence for the fact that she believes that p. Along these lines, Borgoni (2016) and Mandelbaum (2016) argue that we have strong reasons to attribute a belief to an agent if she explicitly endorses its content.[14] What is important for our present discussion is this: with the layer-account, we have crucial resources at our disposal to argue that an agent has a particular valuing even if she lacks the characteristic dispositions. The layer-account can thus assign valuings in cases where the surface-account would rather deny them.

Let us finally explain what the accounts predict for the time when Barak begins to have difficulties in his life and, therefore, has a conversation with Abby. At this point, Barak also sincerely asserts that he values his partnership, but he already lacks several of the relevant dispositions or has them only to a lesser degree. Crucially, however, unlike in Barak's depressed phase, he still shows some willingness to resolve tensions and has other characteristic dispositions that he will later lose.

Hence, similar to the phase of Barak's depression, the layer-account would most likely judge that Barak values his partnership in this phase. After all, not only does Barak sincerely assert that he values his partnership, but he also shows several of the relevant dispositions. These dispositions may serve as additional evidence for Barak's having the respective mental representation. Thus, a proponent of the layer-account has even more reason to believe that Barak values his partnership.

The surface-account would not reach such a clear judgment in Barak's transition phase: the account seems ambiguous about whether we should attribute the valuing to Barak because it is unclear whether Barak has the relevant dispositions *to a sufficient degree*. This is a likely point of disagreement among different proponents of the surface-account, who might disagree over the precise threshold below which an actor's dispositions are too impoverished to constitute valuing. Another source of disagreement among proponents of the surface-account might be whether valuing is a categorical matter or a matter of degree. Therefore, there is some latitude in how a proponent of the surface-account might evaluate Barak's valuing in that phase.

So far, we have introduced the two accounts of valuing and explained their main theoretical commitments. In addition, we have shown that they can reach different verdicts about which valuings agents have, and that in cases where their verdicts converge, they cite different reasons for why valuing obtains. Now, the aim of this paper is not to ultimately adjudicate between the two accounts, i.e., to argue that one of them should be preferred over the other. This is a task that would require careful argumentative effort decidedly beyond what we can reasonably do in this paper. Still, we would like to make some efforts that can at least help soften the ground for future discussions. Towards this aim, we will now embed the distinction within two extant philosophical debates. The first debate – between *representationalists* and *dispositionalists* – concerns the correct conceptualization of belief. We will illustrate that an

[14] Note that while those authors hold that a sincere assertion of the kind "I believe that p" is sufficient evidence for the fact that one believes that p, they do not endorse that believing p can be reduced to a disposition to sincerely assert that p.

understanding of this debate can help with adjudicating between the surface- and the layer-account. The second debate is about different accounts of valuing offered in the meta-ethical literature. As we will show, this debate stands to benefit from an explicit recognition of the distinction we offer here.

3 Surface- and layer-analyses of belief: lessons for the valuing-case

Recently, philosophers of cognitive science have identified and debated two conceptions of belief – *representationalism* and *dispositionalism* – that closely mirror our distinction concerning valuing. Roughly speaking, representationalism views belief essentially as a matter of having a certain representational state, and dispositionalism as a matter of possessing particular dispositions. In this section, we will argue that an understanding of the debate between these two positions can help identify some of the main dialectical fault lines to be expected in a debate between proponents of the surface- and the layer-account, despite also revealing important peculiarities of the valuing-case that make it particularly difficult to navigate. Moreover, having the belief-case in mind is essential for our later discussion, since many meta-ethicists who analyze the concept of valuing hold that certain beliefs are central components of valuing.

3.1 In-Between cases of believing

One of the most central issues that sparks off said debate over belief are so-called "in-between cases of believing." Here is a paradigmatic example.

> **Teacher.** When asked whether boys and girls can be equally good at math, a teacher readily asserts that this is the case. However, with an eye on the teacher's behavior, we can observe that she consistently overestimates the talent of her male students and underestimates the talent of her female students. The teacher even has to force herself to show appreciation for her female students' skills, while such appreciation comes naturally to her in the case of her male students. Now, the crucial question is: does the teacher really believe that boys and girls can be equally good at math?

As Schwitzgebel notes, in such cases there is no clear answer to the question of whether or not the agent has the respective belief. On the one hand, the fact that the teacher asserts that boys and girls are equally good at math is evidence that she has the corresponding belief. Yet, her behavioral and emotional dispositions indicate otherwise. There is thus considerable theoretical uncertainty as to whether the teacher has the respective belief. As Schwitzgebel points out, any viable account of belief should be able to reproduce, rather than smoothe over, the uncertainty present in "in-between cases of believing." Yet, representationalism, which Quilty-Dunn and Mandelbaum (2018, p. 2354) call the "orthodoxy in the philosophy of cognitive science," fails to deliver in this respect, or so Schwitzgebel argues.

According to representationalism, to have a belief is to possess an internally stored mental representation (see Quilty-Dunn and Mandelbaum, 2018). Thus, as Schwitzgebel highlights, under representationalism there is always a clear yes-or-no answer to the question of whether an agent has a certain belief – after all, an agent either possesses or lacks a particular (internally stored) mental representation. Therefore, he concludes that representationalism is unable to explain the uncertainty present in "in-between cases of believing."

Schwitzgebel argues that this supports his favored alternative conception of belief: dispositionalism. According to it, "[t]o believe that p […] is nothing more than to match to an appropriate degree and in appropriate respects the dispositional stereotype for believing that p" (Schwitzgebel, 2002, p. 253). It is important to note that Schwitzgebel's dispositionalism is more nuanced than traditional versions of dispositionalism, which characterize beliefs exclusively in terms of *behavioral* dispositions (see Ryle, 2009; Dennett, 1987). The set of stereotypical dispositions, which Schwitzgebel employs to analyze beliefs, can include a variety of dispositions such as behavioral, cognitive, and phenomenal dispositions.

Dispositionalism can elucidate "in-between cases of believing" because it might turn out that agents possess some of the characteristic dispositions while lacking others. If this is the case, dispositionalism might license the verdict that the agent "kind of does, but also kind of does not" have the belief in question.

But representationalists have tried to accommodate "in-between cases" as well. Quilty-Dunn and Mandelbaum argue that not all mental representations need to be equally well accessible by each of an agent's cognitive systems – a representation "can be accessible more or less easily to a single system" (Quilty-Dunn & Mandelbaum, 2018, p. 2359). In the teacher-case, Quilty-Dunn and Mandelbaum would say that the teacher possesses *two* representations, i.e., two beliefs, each of which plays a role in a different system. First, she may possess the representation that boys and girls can perform equally well at math, and second, the representation that boys are principally better at math than girls. The former may be readily accessible if the teacher is engaged in active deliberation and is thus responsible for her assertions. The latter representation, however, may not be readily accessible in such situations. Yet, it is accessible to the system that guides the teacher's behavior towards her students. Referring to these two distinct representations (and distinct cognitive systems) could allow representationalists to account for "in-between cases of believing."

3.2 The main fault line in the belief-debate

As should be clear, the distinction between representationalism and dispositionalism mirrors the distinction we draw in the valuing-case: the former views belief as a layer-, the latter as a surface-phenomenon. Now, we think that the dispute about the belief-case is quite instructive for future discussions of the valuing-case. To see why, let us consider the kinds of considerations that are put forward in debating the belief-case.

In our view, the crucial determinant in the belief-debate is what theorists want the concept of belief to do. Proponent of representationalism, like Quilty-Dunn and Mandelbaum (2018, p. 2354, p. 2362, p. 2369), frequently highlight that they favor

it because it construes "belief" such that it can be fruitfully integrated into cognitive science explanations. For instance, in the teacher-case, representationalism can *explain* the conflict between the teacher's assertions and her behavior by referring to the agent's two mental representations that produce this conflict. In contrast, dispositionalism merely diagnoses *that* the teacher has mixed dispositions without further explaining this fact via reference to the agent's cognitive makeup. Representationalism thus comes with the promise of increased explanatory depth. It should be noted, however, that in attempting to provide such explanations, representationalism also takes on greater empirical commitments. For the explanation in our example to go through, there must be two distinct cognitive systems that have access to different representations.[15]

Dispositionalists, on the other hand, hold that a main advantage of dispositionalism over representationalism lies in the fact that it takes a decidedly *practical viewpoint*. This viewpoint comes with several merits. For instance, construing belief along the lines of dispositionalism can help navigating the social world. As Schwitzgebel explains, when we think about people's beliefs, we should be "interested in people's general postures toward the world," and our conception of belief should therefore inform us primarily about "how people will act when it matters" (Schwitzgebel, 2021, p. 363). This idea is reminiscent of Dennett (1987), who argues that we should assign to agents those beliefs and preferences that provide us with the most accurate predictions of general trends in their behavior. As dispositionalism is tailored to establish a tight connection between belief and patterns in our acting, thinking, and feeling, Schwitzgebel considers it practically advantageous in virtue of capturing patterns that allow us to successfully navigate the social world.[16]

Another advantage concerns the evaluative function of belief: dispositionalism, Schwitzgebel (2021, pp. 363–364) argues, encourages thorough self-examination. Specifically, he points out that under dispositionalism we can only be certain about whether we really believe p if we have first engaged in a thorough self-examination of how we act, think, and feel. Otherwise, we cannot be sure whether or not we have the dispositions associated with believing p. So, it is not sufficient to sincerely assert that one believes p in order to be sure that one really believes p (whereas, under representationalism, such an assertion seems sufficient). The fact that dispositionalism encourages thorough self-examination thus precludes the possibility of providing spurious excuses.

In light of this overview, we conclude that the main fault line in the debate on belief – at least as it is currently conducted – is not primarily empirical or an issue of conceptual argument. At its heart, it is a matter of what practical ambitions we have regarding the concept of belief; in short, it is a matter of what we want the concept of belief to do (e.g., figuring into cognitive science explanations vs. encouraging thorough self-examination).

[15] Dispositionalists like Schwitzgebel object that this explanation does not stand up to empirical scrutiny.

[16] Note that this connection would be somewhat obscured by conceptions of dispositions other than the simple conditional analysis that would ascribe dispositions even in the presence of finks or masks (e.g., Manley and Wasserman, 2007; Vetter, 2015). Again, while we do not wish to say that these conceptions have no appropriate contexts of applications, we here rely on the simple conditional analysis to outline the surface-account (and dispositionalism).

3.3 Lessons for debating the valuing-case

We hold that the main fault line in the debate about beliefs carries over to debating the two accounts of valuing we put forth in this paper. Also in the valuing-case, we think that considerations having to do with our ambitions for the concept of valuing will be relevant in adjudicating between the surface- and the layer-account. For example, we think that proponents of the layer-account will tend to stress that their account locates valuing within an agent's mental architecture and can thus facilitate "deeper" explanations of the agent's behavior. In contrast, proponents of the surface-account might emphasize that their account can better illuminate "people's general postures toward the world" as manifested in their everyday actions, emotions, and deliberations. Moreover, the surface-account, like dispositionalism about belief, encourages thorough self-examination because it invites us to confirm that we really exemplify the dispositions associated with a particular type of valuing. Hence, depending on one's ambitions concerning the concept of valuing, one has reasons to favor one account over the other.

Note, however, that some of these ambition-related, pragmatic considerations may ultimately depend on our own valuings. For instance, is thorough self-examination really important to us? This generates a considerable difficulty in debating the valuing-case, that is absent in the belief-case. For in the belief-case, we could in principle switch back and forth between two different understandings of belief depending on which pragmatic goal we are pursuing. Yet, when it comes to deciding which goals we should pursue and which priorities we should attach to them, what we value seems to become relevant. This seems to introduce an additional difficulty for deciding between the surface- and layer-account. To give a simple illustration of the problem, consider that one favors the surface-account as it promotes thorough self-examination. Yet, applying the surface-account to oneself may lead one to realize that one does not really value thorough self-examination. One only asserts that one does. Hence, endorsing the surface-account can exert pressure against one's endorsement of the pragmatic goal that led one to adopt the account in the first place. We, therefore, think that the fact that adjudicating between the accounts is dependent on our own valuings introduces another layer of complexity to the debate that is absent from the belief-case.[17]

Having pointed out some of the aspects that we think will, and should, play a role in assessing the accounts, let us reiterate that the main aim of this paper is not to argue for adopting one over the other. Such a discussion will have to wait for another occa-

[17] Of course, if one wants to adjudicate between representationalism and dispositionalism about belief (without retaining the latitude to switch back and forth between them depending on one's purpose), one must also first settle the question of what one values. This is a complexity inherent to both the belief- and the valuing-debate. The difficulty specific to the valuing-debate, however, has to do with the fact that the question of whether one opts for the surface- or the layer-account is not independent of the preceding question just mentioned: in committing to one account or the other one ultimately also takes a stance on what one values.

sion, and it will probably involve more – and perhaps even quite substantial – points of contention than we have been able to anticipate based on the belief-case.[18]

Rather, we will now illustrate the urgent need to appreciate and philosophically engage the distinction between the surface- and the layer-account in the first place. In particular, we will situate the distinction within the current meta-ethical literature on valuing with the aim of showing that it offers little to adjudicate between these accounts. Indeed, in light of the distinction we have proposed, supposed fault lines within the current dialectic disappear upon closer examination, and we are left with the question of whether the surface- or the layer-account is true. This not only reveals an implicit disregard of the proposed distinction and thus the need to engage it philosophically. It also reinforces the suspicion that the distinction we propose reflects a profound way of discerning competing conceptions of valuing.

4 The two accounts and the meta-ethical debate on valuing

Within the contemporary philosophical debate on how to analyze the phenomenon of valuing, two camps have evolved: first, philosophers who identify valuing with the presence of a certain propositional attitude, and second, philosophers who emphasize the importance of dispositions involved in valuing. At first sight, one might think that the debate already reflects the distinction that we offer in this paper. On closer inspection, however, it turns out that current accounts are more or less ambiguous vis-a-vis their acceptance of either the surface- or the layer-account. Moreover, when the proposed distinction is brought to the fore, it becomes apparent that scholars may talk past each other within the current dialectic, with the question still lingering whether to opt for the surface- or the layer-account. In what follows, we will consider both camps in due course, highlight the aforementioned ambiguity, and finally draw some lessons for the extant dialectical situation.

4.1 Propositional attitude theories

Philosophers of the first camp propose that we can analyze valuing in terms of the presence of a certain propositional attitude that is either cognitive or conative. In this manner, Gary Watson states that "valuing is essentially related to thinking or judging good" (Watson, 1975, p. 208). Lewis, who draws on Harry Frankfurt's hierarchical model of agents' desires, suggests that "valuing is just desiring to desire" (Lewis, 2000, p. 71).[19] Further, Smith (1992) analyzes valuing as the belief that the thing in question is valuable, which means that one has particular normative reasons to act in certain ways with regard to it.[20]

[18] Let us emphasize that a discussion of the two accounts may also reveal that the correct view is pluralism, i.e., the view that both accounts are equally valid, especially as they may have their raison d'être in different contexts. We thank an anonymous reviewer for drawing our attention to the possibility of pluralism.

[19] See Taylor (1985) for a related proposal.

[20] For a related proposal, see Scanlon (1998, p. 95). As a further propositional attitude theorist Michael Bratman comes to mind. He develops an account according to which valuing is to be analyzed not by a single mental state but by a combination of two states; he tells us that "an agent values X (in the relevant

As far as we know, philosophers who advocate such *propositional attitude theories* do not explicitly address the question of whether the propositional attitude that is supposed to be constitutive of valuing should be interpreted along the lines of representationalism or dispositionalism.[21] It certainly seems the more natural interpretation to read these accounts as being implicitly based on representationalism. After all, representationalism is the "orthodox view" about the nature of mental states (including propositional attitudes). Hence, their positions lean rather toward endorsing the layer-account. Indeed, some remarks by Watson are indicative of the layer-account, for example, when he explains that "an agent's values[, i.e., valuings] consist in those principles and ends which he – in a cool and non-self-deceptive moment – articulates as definitive of the good, fulfilling and defensible life" (Watson, 1975, p. 215). Such reflection on one's conative representations would indeed be a good guide to discovering one's valuings under the layer-account.[22] However, propositional attitude theories can in principle also be interpreted as surface-accounts if we subscribe to dispositionalism about mental states.[23] The fact that most proponents of propositional attitude theories may exhibit a closer proximity to the layer-account in their writings does not detract from this possibility.[24]

sense) when it has a desire for X […] and a self-governing policy in favor of treating that desire as providing an end that is justifying (perhaps to a certain, specified degree) in motivationally effective deliberation" (Bratman, 2007, p. 65).

[21] Note that we talk of "theories" and not mere "conceptions" of valuing here. We consider theories of valuing as being more extensive than mere conceptions: theories of valuing, as set forth in the literature, not only offer a particular notion of what valuing is, but they also provide theoretical justifications supporting that notion. In other words, theories of valuing include a rich body of theoretical resources (like, for instance, arguments) and, as an end-result, provide a particular notion – or conception – of valuing. Lewis, for example, not only puts forward the conception that valuing is desiring to desire, but with it also a set of theoretical arguments favoring this conception (see Lewis, 2000, pp. 69–73). We thus think that the relation between theories of valuing and conceptions of valuing is roughly analogous to how Rawls (1971) sees the relation between his theory of justice and the conception of justice as fairness, where the latter is the theory's end-product.

[22] See also Williams (1973, p. 118) and Frankfurt (1982, p. 260).

[23] To be sure, the type of dispositionalism would have to be what we refer to in footnote 10 as thick-dispositionalism rather than thin-dispositionalism.

[24] Recently, Heathwood (2019) has put forth an interesting distinction between two kinds of desires: "behavioral desires" and "desires in the genuine-attraction sense." Roughly speaking, "behavioral desires" are those that reflect what the agent wants "in a merely behavioral sense, in that the person is, for some reason or other, disposed to act so as to try to get it," and "desires in the genuine-attraction sense" are those reflecting "what a person wants in a more robust sense, the sense of being genuinely attracted to the thing" that goes along with "with notions like enthusiasm" (Heathwood, 2019, p. 664, p. 673). Importantly for Heathwood, the difference here is not merely phenomenological, but concerns "the way of relating to the object of the desire" (Heathwood, 2019, p. 674). As Heathwood argues, only desires in the genuine-attraction sense are relevant for *wellbeing* – thus, the question arises whether one could also regard them as relevant for *valuing*. Specifically, the question emerges how they relate to the surface- and the layer-account. Quite generally, we think that desires in the genuine-attraction sense are compatible with either the surface- and the layer-account. Along the lines of the layer-account, desires in the genuine-attraction sense could be conceptualized as representational mental states with a particular intentional aspect. Along the lines of the surface-account, one could conceptualize desires in the genuine-attraction sense as a set of dispositions. Yet, crucially, these dispositions would have to be restricted to emotional and deliberative dispositions and not include behavioral dispositions because the latter do not bear a constitutive relation to genuine-attraction desires. While the absence of those behavioral dispositions would make valuing – based on genuine-attraction desires – different from how we characterize valuing under the surface-account here,

4.2 Multiple feature theories

Let us now turn to the second camp, where the ambiguity is much more pronounced. We call this class of accounts *multiple feature theories*. The most influential philosopher here is arguably Scheffler (1997, 2004, 2011), whose conception of valuing has been central to the work of scholars such as Kolodny (2003) and Wallace (2013). These philosophers have criticized propositional attitude theories for not including dispositions in their accounts.[25] In particular, Scheffler (2011, pp. 33–34) criticizes propositional attitude theories for not "identify[ing] emotional vulnerability as an aspect of valuing."[26] In response, he develops an account that emphasizes the crucial role of dispositions. In some passages, Scheffler even seems to go so far as to characterize valuing merely by reference to an agent's dispositions, for instance, when he explains that valuing something is "a distinctive way of being favorably disposed toward it" (Scheffler, 2011, p. 30).[27] This seems to fit well with the surface-account.

However, in spite of that, Scheffler frequently states that valuing also involves a certain belief, namely, a belief that the thing in question is valuable. In a summary of his account, Scheffler (2011, p. 32) proposes that "valuing any X involves at least the following elements:

1) A belief that X is good or valuable or worthy,
2) A susceptibility to experience a range of context-dependent emotions regarding X,
3) A disposition to experience these emotions as being merited or appropriate,
4) A disposition to treat certain kinds of X-related considerations as reasons for action in relevant deliberative contexts."

In our view, Scheffler's account does not show an unequivocal commitment to either the surface- or the layer-account. First of all, note that his account merely states that valuing "involves" the listed elements. Thus, it seems unclear whether the listed elements are intended to be constituents of valuing or merely typical effects of valuing, and so whether he proposes a constitutive account of valuing.[28] One consequence of this theoretical lacuna is that the account leaves open how we should treat the individual elements. It may be that some of the listed features are essential to valuing, while others are merely co-occurring with valuing. In particular, it might turn out that only the listed dispositions (see conditions 2–4) are essential to valuing, whereas the belief state (see condition 1) is just frequently co-occurring in agents who have a cer-

this fact should not dissuade one from further investigating the possibility of developing a surface-account based on genuine-attraction desires. We thank an anonymous reviewer for highlighting the need to engage with Heathwood's insightful article.

[25] See also Anderson (1993).

[26] In the passage cited, Scheffler is particularly critical of Scanlon's account, which is very similar to Smith's. Yet, Scheffler holds that *all* propositional attitude theories on the market are defective because they do not make reference to dispositions such as emotional vulnerability.

[27] See also Wallace (2013, p. 23).

[28] Note that both Kolodny and Wallace adopt Scheffler's talk of 'involve' and thus equally seem to refrain from offering a constitutive account (see Kolodny, 2003, p. 150; Wallace, 2013, p. 23).

tain valuing. In fact, given the centrality Scheffler ascribes to dispositions – not only as crucial components of valuing, but also as marking the main dialectical contrast to propositional attitude theories – the purpose of the belief state in Scheffler's account remains somewhat opaque.

This opacity allows us to construe Scheffler as a proponent either of the layer- or the surface-account. As we have already noted with respect to propositional attitude theories, there is always the theoretical possibility to interpret the invoked belief state in dispositional terms. This possibility is particularly alive concerning the belief state involved in Scheffler's account. First, if one fleshes out the "belief that X is valuable" as the belief that one has reasons to act in certain ways with regard to it, then this belief comes very close to the fourth condition; and it can then be argued that said belief is nothing but the disposition to "treat certain kinds of X-related considerations as reasons for action." What is more, Scheffler himself brings into play another possible view for analyzing this belief, according to which "the belief that X is valuable just is, in part, the belief that one's context-dependent emotional reactions regarding X are merited." He goes on and explains that, under this analysis, "it will seem a redundant feature of my account that it includes, as separate items, both the belief that X is good or valuable and the disposition to experience one's context-dependent emotions regarding X as being merited or appropriate." It is noteworthy that, here, Scheffler seems to be fine with analyzing the relevant belief in terms of dispositions. And although he is reluctant to eliminate the first condition from his account – on grounds of independent meta-ethical considerations[29] –, he seems quite relaxed about this possibility. For those of us who do not share his other meta-ethical convictions, he concedes that "the redundancy in [his] account is easily remedied," i.e., by way of "simply eliminat[ing] the first element in [his] account of valuing, while retaining the other three" (Scheffler, 2011, p. 33, p. 42 n. 38). Scheffler thus seems to think that the core of his account would remain intact even in this scenario. Finally, we hold that there is also a some systematic pushback to the idea that the respective belief is a mental representation: conceiving of all conditions in dispositional terms would lead to a theoretically more unified and parsimonious account compared to defining valuing in terms of an ontologically diverse set of entities (that includes mental representations as well as dispositions).[30] Hence, there are good reasons for construing Scheffler as a proponent of the surface-account.

Nevertheless, Scheffler's view might also be considered a version of the layer-account. There are at least two ways of reading him in this fashion. First, recall that his account is surrounded by considerable theoretical uncertainty since Scheffler merely claims that valuing "involves" the listed features. Consequently, the account also leaves open the possibility that there is more to valuing besides these features. For example, even if all the listed conditions describe dispositions – and not representational states –, it could be that there is an underlying mental representation that

[29] In particular, he argues that this issue hinges on whether or not one accepts what he calls the *priority claim*, according to which "valuing is a notion that is prior to believing valuable, and that believing valuable is to be explained in terms of valuing" (Scheffler, 2011, p. 36).

[30] See also Knobe and Preston-Roedder (2009) for another seemingly mixed view, referring to both mental states and dispositions.

causes these dispositions; or, even in case the belief mentioned in the first condition turns out to be a representational state, it might still be that there exists a further underlying representational state that implies the "complex syndrome of interrelated dispositions and attitudes" (Scheffler, 2011, p. 32). In both these scenarios, we would have instantiations of the layer-account.

Second, suppose that the belief-state in the first condition in fact constitutes a representational mental state. Then, it could be that this very mental state causes the obtainment of the dispositions formulated in the other conditions. Surprisingly, Scheffler at times suggests that this might be a genuine possibility: "it may be said that even believing something valuable *brings with it* some degree of emotional vulnerability" (Scheffler, 2011, p. 42 n. 39; emphasis added). In sum, Scheffler's influential account, that gives special importance to dispositions, can be read as a surface- or a layer-account.

4.3 Implications for the extant dialectic

As we saw, the proposed distinction has not been sufficiently acknowledged in the philosophical debate on the nature of valuing. Yet, explicitly recognizing it could benefit the debate: because the distinction differentiates two opposing views about what valuings are (representations or dispositions), it promises to help theorists formulate and develop their accounts in ways that get clear on this fundamental dimension and thereby eliminate the highlighted ambiguities.

Conversely, without an awareness of the distinction, theorists run the danger of talking past each other. This danger, we believe, already materializes in practice. Reconsider Scheffler's main argument against propositional attitude theories: that they do not include dispositions at all and that this leads to an impoverished view of valuing. As he notes dismissively, "only the temptations of bad theory lead people occasionally to find ways of forgetting it" (Scheffler, 2004, p. 254). But as should be clear by now, proponents of the layer-account need not forget or deny that dispositions play an important role in valuing. After all, propositional attitude theorists, who subscribe to the layer-account, can acknowledge that dispositions are crucial effects of valuing, whereas they would insist that valuings are, at bottom, a matter of the presence of a certain representational state. They might therefore see no problem in Scheffler's critique because it is based on a spurious dichotomy. Yet, they could still complain that he overstates the importance of dispositions by invoking them as essential theoretical elements. A constitutive account, they might argue, would not explicitly include them. Alternatively, proponents of propositional attitude theories could spell out what it means to have the relevant propositional attitudes in dispositional terms and thus arrive at a version of the surface-account. Pace Scheffler, this would secure dispositions a central role in their analysis of valuing. Overall, this shows that an awareness of the distinction brings to light that supposed fault lines that exist in the current debate on valuing actually vanish on closer inspection and effectively turn into the very issue of whether one conceives of valuing along the lines of the surface- or the layer-account.

Against this background, it seems that the distinction between the surface- and the layer-account marks a fundamental divide in how to think about valuing. One

diagnosis for why this might be so is that the surface- and the layer-account represent two contrasting and salient notions of *what it is* to value something, i.e., they answer the question of what kind of entity constitutes our valuings (representational mental states or dispositions). As such, these accounts speak to a core ontological issue with respect to valuing, whereas the extant meta-ethical debate, as we saw, is largely silent on this matter.[31]

Because the accounts address the question of *what it is* to value something, they lend themselves to informing and advancing fields of practical research that implicitly depend on an answer to this question. Specifically, areas of social scientific research come to mind that treat valuing as a central measurand.[32] In the remainder of the paper, we will demonstrate the theoretical leverage of the proposed distinction to such areas of practical research, using the example of behavioral welfare economics.

5 The distinction's theoretical value for behavioral welfare economics

We hold that acknowledging the distinction between the surface- and the layer-account has at least two noteworthy ramifications for behavioral welfare economics. First, it helps uncover crucial meta-normative assumptions in this field: a close examination reveals that most approaches to measuring wellbeing in behavioral welfare economics are implicitly committed to the layer-account. Second, this result also brings to light the theoretical possibility of understanding and conducting behavioral welfare economics based on a surface-conception of valuing.

Let us start with some basics. Economics still relies predominantly on *preference satisfaction* as a welfare criterion (see Adler and Posner, 2006; Thaler and Sunstein, 2008).[33] It is standardly assumed that an agent's preferences are closely related to her wellbeing if her preferences satisfy certain criteria: transitivity, completeness, stability, and context independence. Now, the basic idea of behavioral welfare economics (for short, BWE) is that an agent's *actual preferences* usually violate these criteria and are therefore not a good guide to wellbeing. Consequently, the thinking goes, we should not be concerned with agents' preferences as they actually manifest themselves in choice behavior. Instead, we should rely on a purer – or laundered – set of preferences. BWE scholars assume that these purified preferences reflect an agent's valuings and are thus closely tied to her wellbeing. Of course, the aim of BWE to launder or purify preferences can appear quite peculiar to outsiders. However, preference-based measures of welfare are quite important for many high-stakes scenarios. Just think of integrated assessment models in the context of climate change and their reliance on GDP (see Weitzman, 2007; Nordhaus, 2018). If preferences are unfit to

[31] Of course, it may be that the authors in this debate *intend* their theories to speak to the ontological question that our proposed distinction throws light on. But even then, our point remains that these theories do not take a clear position on this question and therefore do not address it in a relevant sense.

[32] See also Lohse (2017), who emphasizes the need to pay attention to the ontological assumptions within social scientific theories, including those implied by their fundamental concepts.

[33] We use the terms "welfare" and "wellbeing" interchangeably.

reliably track our welfare, many of these models, which are meant to inform concrete policy decisions, would be in serious trouble.[34] Hence, we take the aim of BWE to rescue preference-based measures of welfare to be highly relevant.[35]

As we now turn to show, large sections of BWE are implicitly committed to the layer-account of valuing. Crucial evidence for this claim comes from a recent and highly influential paper by Gerardo Infante, Guilhem Lecouteux, and Robert Sugden (2016, henceforth ILS). In this work, the authors unfold what they take to be the "new consensus" view in BWE, and in particular its ontological assumptions about what purified preferences are. Specifically, they aim to show that the idea of preference purification presupposes a dualistic model of the mind, according to which a so-called inner rational agent is trapped inside a psychological shell.

To see that this model involves a commitment to the layer-account, let us explain ILS' reconstruction of BWE in more detail. ILS hold that the only way to rationalize preference purification is by attributing a certain set of cognitive mechanisms to individual agents. First, we must attribute a mechanism that generates *pure* preferences, i.e., preferences that satisfy properties like transitivity, completeness, stability, and context independence. ILS call this mechanism the *inner rational agent*. Second, according to ILS, BWE scholars must further assume the existence of a mechanism that distorts the preferences generated by the inner rational agent. This yields the agent's actual preferences, which are usually messy, i.e., they do not satisfy the just mentioned properties. This second mechanism, that distorts the agent's pure, or "deep down," preferences, is called the *psychological shell*. In light of this reconstruction, ILS centrally claim the following: preference purification should be understood as the attempt to recover the output of the inner rational agent from the distortions produced by the psychological shell.

This reconstruction naturally lends itself to the layer-account. After all, according to ILS, the agent's valuings are identified with the "deep down" preferences produced by the inner rational agent. Moreover, by our lights, these preferences must be viewed as representational states. The reason for this is that the outputs produced by the inner rational agent are thought to be available for further computational processing by the psychological shell, which leads to the agent's (messy) preferences that actually govern her choice behavior. For this story to make sense, preferences must be the kinds of entities that can serve as inputs and outputs of specific psychological processes like the inner rational agent and the psychological shell. This means that we need to understand the "pure" preferences relevant for valuing as representational states.

[34] Hence, in contrast to older debates about cost-benefit-analysis, which focused on the issue of ignoring all signs of people's values beyond what they express in their choices (compare this to what we call thin-dispositionalism in footnote 10), the current debate in BWE starts with the recognition that our actual choices can be bad guides to what we truly value.

[35] Note that the term "valuing" is not explicitly used by all authors in this literature. Instead, we often read, for instance, of people's "subjective interests" (e.g., Thoma, 2021). However, for our purpose, we hold that terms such as "subjective interest" and "valuing" mean the same. We take it for granted that those representational states that are supposed to ground "subjective interests" according to authors like Thoma are the same states that ground valuing under the layer-account. Moreover, other authors like DesRoches (2020) explicitly speak of "values-based preferences" to denote the distinct type of attitude that BWE scholars are (allegedly) trying to uncover.

The only other option in the present dialectical context would be to construe the output of the inner rational agent as a complex set of dispositions. To see that this is not plausible, first note that these dispositions are likely to have highly distributed causal bases that are not easily localized. Hence, why should we expect that they are the output of a single psychological mechanism, i.e., the inner rational agent? Unless, of course, we assume that all those dispositions ultimately depend on the presence of a particular representational state, which in turn is the output of the inner rational agent. But that would again collapse into the layer-account. What is even more problematic about construing the output of the inner rational agent as the complex set of dispositions identified by the surface-account is that those dispositions are agent-level dispositions instead of dispositions of any sub-system like the inner rational agent. If the agent were to already exhibit the relevant (agent-level) dispositions, it would become unclear why we even need to postulate the inner rational agent to begin with: according to the surface-account, an agent's valuing would simply be identified with the agent's dispositions. Any fact about the precise psychological mechanisms on which these dispositions depend would be inessential for identifying an agent's valuings. As a result, determining the output of the inner rational agent would be irrelevant for identifying what the agent values. All in all, then, ILS' reconstruction of BWE suggests that it is inherently linked to the layer-account.

ILS believe that their reconstruction yields a strong argument against the very aim of BWE, which consists in developing measures of wellbeing that are based on agents' preferences while admitting situational influences on their choices. The reason for ILS' skepticism is that BWE – with its commitment to a dualistic picture of the mind – rests on a highly implausible picture of human psychology. In particular, ILS hold that our best psychological theories do not support the picture of an inner rational agent whose output gets distorted by a psychological shell.[36]

In the face of this criticism by ILS, advocates of BWE have put forth various replies. A recent example is Thoma (2021), who provides a reconstruction of BWE that does not require inner rational agents, or so she argues.[37] According to Thoma, we should neither regard preferences as fundamental nor identify them with valuings. Instead, valuings should be seen as "constituted by more fundamental attitudes, call them desires, on the basis of which preferences are typically formed." Thus, she holds that the correct, "new normative foundation for welfare measurement [is essentially] 'desire-based'" (Thoma, 2021, pp. 356–357). Thoma's view rests on the idea that when preferences satisfy the characteristic formal properties that economists standardly require them to satisfy, we can be confident that they resulted from a uniquely correct aggregation of our fundamental desires.[38] However, if our fun-

[36] One could argue that inner rational agents should not be understood as real psychological entities within the agent but rather as an idealized version. Yet, according to ILS, BWE needs to rely on inner rational agents as psychological entities because only this will allow BWE to claim that they still judge the agents according to their own lights (for further discussion of this argument, see Beck, 2022).

[37] While it does not matter much for our argument, Thoma does not think that her reconstruction works for all methods of preference purification. Instead, she only focuses on the work of Bernheim and Rangel (2008; see also Bernheim, 2016).

[38] For a distinction similar to Thoma's desire-preference distinction, see Heathwood's (2019) distinction between desires in the genuine-attraction sense and behavioral desires.

damental desires are not precise enough (or too vague) to be uniquely aggregated into preferences, Thoma (2021, p. 360) holds that contextual influences can impact our preferences. According to Thoma, such contextual influences are evidenced by the fact that preferences do not satisfy the formal properties that economists usually require them to satisfy, e.g., different permissible aggregations of vague desires can lead to intransitive preferences. Moreover, the impact of contextual influences on our preferences also means that preferences cannot deliver a determinate welfare ranking. Yet, according to Thoma, in cases where preferences result from a uniquely correct aggregation of our desires, we can still build preference-based measures that deliver a determinate ranking of options. Moreover, even in cases where preferences cannot be uniquely aggregated, we can, at least sometimes, build incomplete welfare measures and "live with this indeterminacy" (Thoma, 2021, p. 260). Thoma concludes that none of this requires us to commit to inner rational agents that always generate preferences that meet all the relevant properties.

Whether or not one considers Thoma's response successful, the following observation is crucial for our purposes: Thoma, like ILS, reconstructs BWE by essentially positing a mechanism that takes certain mental states as inputs (i.e., desires) and produces other mental states (i.e., preferences) as outputs. Thus, for the same reasons as ILS' reconstruction of preference purification, Thoma's account is tied to a representational state interpretation of the desires that, according to her, ground valuing.[39] This suggests that her account, too, is committed to a layer- rather than a surface-account of valuing.[40]

The fact that BWE – as it is construed by critics and proponents alike – involves a layer-understanding of valuing reveals a crucial implicit meta-normative assumption behind much of the current work in this area. Let us highlight two implications of this. On the one hand, if you are a proponent of the surface-account, you will not find the current practice of BWE particularly appealing. This holds for both Thoma's positive and ILS' negative reconstruction. More specifically, with respect to ILS' reconstruction, you would wonder why the output of the inner rational agent is taken to have normative relevance in the first place. Similarly, although Thoma's view promises to dispense with the inner rational agent, it still would not accomplish much

[39] One may argue that Thoma could accept a dispositional approach to desire but treat contextual influences as distorting factors or masks. However, we are skeptical that this is a plausible construal of her account. First, the fact that Thoma's account works with multiple desires, which can relate to the same options, makes an interpretation along the surface-account difficult. After all, the surface-account works with agent-level dispositions. But how could we understand these various desires as agent-level dispositions (as opposed to sub-agential functionally (or dispositionally) individuated mental states)? Second, suppose we understand desires here as being associated with agent-level dispositions. In this case, it is unclear what it would mean to aggregate these various desires, as suggested by Thoma's account. Third, even if there is only one relevant desire, but this desire is distorted by contextual influences, how do we differentiate the context in which the agent manifests her true dispositions from those in which she fails to do so? We suspect that such a distinction is difficult to uphold without stipulating something that will look suspiciously like a layer-account. Hence, we think that Thoma is most plausibly construed as a layer-theorist. We thank an anonymous reviewer for raising this issue.

[40] See also DesRoches (2020, p. 566) for further evidence that most scholars working on BWE portray it as relying on the layer-account.

if one adheres to the surface-account. After all, even on her view, valuing would still be identified with some kind of representational state.

On the other hand, the surface-account points to an interesting theoretical avenue for proponents of BWE, i.e., they could premise their endeavors on a surface-account of valuing. In this regard, it is worth noting that there are indeed economists who seem to commit to a surface-account and yet hold that (certain forms of) preference purification make sense (see, e.g., Harrison and Ross, 2018). Those authors advocate Dennett's (1987) intentional stance, which we think ultimately commits them to a version of the surface-account. (Dennett's view exclusively refers to *behavioral* dispositions and would, therefore, provide a rather sparse version of the surface-account).

On top of that, in a response to ILS, Hausman (2016; see also Beck, 2022) advocates a coherence-based understanding of preference purification. According to this view, preference purification should be about asking what an agent would value if she were (more) coherent. This proposal could be reconciled with the surface-account as follows: under the surface-account, there is the possibility of so-called "in-between cases of valuings", i.e., cases in which a look at the agent's dispositional profile does not allow for clear judgments about whether or not the agent actually has a particular valuing. This may include cases in which most of the agent's dispositions suggest that she has a particular value, while some of her *behavioral* dispositions, i.e., her choices, suggest the opposite. Now, an alternative understanding of preference purification that is in line with Hausman's suggestion as well as the surface-account could essentially claim that preference purification is about finding a set of preferences (understood as behavioral dispositions) that matches the valuings one should assign to her based on her overall dispositional profile. In other words, preference purification should not be seen as recovering certain "deep down" representational states, but about finding out how the agent would behave if her dispositional profile perfectly matched the valuings that we assign to her based on the overall dispositions she already has.

In sum, recognizing the distinction between the surface- and layer-account has allowed us to show that much of the current debate on preference purification is based on the layer-account. This not only exposes an important meta-normative assumption in this debate, but also points to an alternative way of understanding and perhaps even conducting BWE.

6 Conclusion

In this paper, we explicated an underappreciated distinction between two accounts of valuing and argued that it is important to bring this distinction to the fore. We began by motivating the distinction between the surface- and the layer-account using the case of Barak. After a detailed explanation of the accounts, we related them to the recent philosophical literature. First, we showed that recent work in philosophy of cognitive science provides useful lessons for discussing the distinction. Second, we argued that debates in meta-ethics stand to benefit from an explicit recognition of the distinction. In the final part of the paper, we suggested that acknowledging the

distinction holds theoretical value for behavioral welfare economics, which depends on a precise understanding of what it is to value something.

Our work opens up countless avenues for future research. Let us mention three of them. First, of course, a detailed evaluation of the accounts is still pending, including an assessment of which of the two accounts is preferable. Here, it would be particularly interesting to see how contemporary valuing theorists relate themselves to the distinction. Second, are there other attitudes besides valuing (and belief) that are worth conceptualizing in either surface- or layer-terms? In this regard, the question arises whether one should be a surface- or a layer-theorist with respect to *all* attitudes, or whether some attitudes are to be conceptualized as surface- and others as layer-phenomena. Finally, beyond the areas mentioned above, are there other areas in which our proposed distinction can provide theoretical leverage? Given the central role of valuing for much of our thinking, we believe that this is the case.

Acknowledgements We would like to thank Anna Alexandrova, Cristian Larroulet Philippi, Isaac Kean, Bele Wollesen, two anonymous reviewers, and the audience at the 11th congress of the Society for Analytic Philosophy (GAP) for their insightful comments and suggestions. Lukas Beck would also like to thank FORMAS for funding his postdoctoral research.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests The authors declare none.

References

Adler, M. D., & Posner, E. (2006). *New foundations of cost-benefit analysis*. Harvard University Press.

Anderson, E. (1993). *Value in ethics and economics*. Harvard University Press.

Beck, L. (2022). *The econ within or the econ above? On the plausibility of preference purification* (pp. 1–23). Economics & Philosophy.

Bernheim, B. D. (2016). The good, the bad, and the ugly: A unified approach to behavioral welfare economics. *Journal of Benefit-Cost Analysis*, *7*(1), 12–68.

Bernheim, B. D., & Rangel, A. (2008). Choice-theoretic foundations for behavioral welfare economics. In A. Caplin, & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook*. Oxford University Press.

Bird, A. (1998). Dispositions and antidotes. *The Philosophical Quarterly*, *48*(191), 227–234.

Borgoni, C. (2016). Dissonance and irrationality: A criticism of the in-between account of dissonance cases. *Pacific Philosophical Quarterly*, *97*(1), 48–57.

Bratman, M. (2007). Valuing and the will. *Structures of agency: Essays*. Oxford University Press.

Choi, S. (2006). The simple vs. reformed conditional analysis of dispositions. *Synthese*, *148*(2), 369–379.
Choi, S., & Fara, M. (2021). Dispositions. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), URL = https://plato.stanford.edu/archives/spr2021/entries/dispositions/
Dennett, D. C. (1987). *The intentional stance*. MIT Press.
DesRoches, C. T. (2020). Value commitment, resolute choice, and the normative foundations of behavioural welfare economics. *Journal of Applied Philosophy*, *37*(4), 562–577.
Frankfurt, H. (1982). The importance of what we care about. *Synthese*, *53*(2), 257–272.
Harman, G. (2000). *Explaining value and other essays in moral philosophy*. Oxford University Press.
Harrison, G. W., & Ross, D. (2018). Varieties of paternalism and the heterogeneity of utility structures. *Journal of Economic Methodology*, *25*(1), 42–67.
Hausman, D. M. (2016). On the econ within. *Journal of Economic Methodology*, *23*(1), 26–32.
Heathwood, C. (2019). Which desires are relevant to well-being? *Noûs*, *53*(3), 664–688.
Infante, G., Lecouteux, G., & Sugden, R. (2016). Preference purification and the inner rational agent: A critique of the conventional wisdom of behavioural welfare economics. *Journal of Economic Methodology*, *23*(1), 1–25.
Knobe, J., & Preston-Roedder, E. (2009). The ordinary concept of valuing. *Philosophical Issues*, *19*(1), 131–147.
Kolodny, N. (2003). Love as valuing a relationship. *The Philosophical Review*, *112*(2), 135–189.
Lewis, D. (2000). Dispositional theories of value. *Papers in Ethics and Social Philosophy*. Cambridge University Press.
Lohse, S. (2017). Pragmatism, ontology, and philosophy of the social sciences in practice. *Philosophy of the Social Sciences*, *47*(1), 3–27.
Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658.
Manley, D., & Wasserman, R. (2007). A gradable approach to dispositions. *The Philosophical Quarterly*, *57*(226), 68–75.
Manley, D., & Wasserman, R. (2008). On linking dispositions and conditionals. *Mind*, *117*(465), 59–84.
McLaughlin, B. P. (2007). Mental causation and Shoemaker-realization. *Erkenntnis*, *67*(2), 149–172.
Nordhaus, W. (2018). Evolution of modeling of the economics of global warming: Changes in the DICE Model, 1992–2017. *Climatic Change*, *148*(4), 623–640.
Pitt, D. (2022). Mental representation. In E. N. Zalta and U. Nodelman, eds., *The Stanford Encyclopedia of Philosophy (*Fall 2022 Edition), https://plato.stanford.edu/archives/fall2022/entries/mental-representation/
Prior, E. (1985). *Dispositions*. Aberdeen University Press.
Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, *175*(9), 2353–2372.
Ramsey, W. (2016). Untangling two questions about mental representation. *New Ideas in Psychology*, *40*(Part A), 3–12.
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Ryle, G. (2009). *The concept of mind*. Routledge.
Scanlon, T. (1998). *What we owe to each other*. Harvard University Press.
Scheffler, S. (1997). Relationships and responsibilities. *Philosophy and Public Affairs*, *26*(3), 189–209.
Scheffler, S. (2004). Projects, Relationships, and reasons. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*. Clarendon.
Scheffler, S. (2011). Valuing. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and recognition: Essays on the philosophy of T. M. Scanlon*. Oxford University Press.
Schroeder, T. (2020). Desire. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), https://plato.stanford.edu/archives/sum2020/entries/desire/
Schulz, A. W. (2018). *Efficient cognition: The evolution of representational decision making*. MIT Press.
Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly*, *51*(202), 76–82.
Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*(4), 531–553.
Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New Essays on Belief*. Palgrave Macmillan.
Schwitzgebel, E. (2021). The pragmatic metaphysics of belief. In C. Borgoni, D. Kindermann, & A. Onofri (Eds.), *The fragmented mind*. Oxford University Press.
Schwitzgebel, E., & Phenomenal, A. (2002). Dispositional account of belief. *Noûs*, *36*(2), 249–275.

Smith, M. (1992). Valuing: Desiring or believing? In D. Charles, & K. Lennon (Eds.), *Reduction, explanation, and realism*. Oxford University Press.
Spurrett, D. (2021). The descent of preferences. *The British Journal for the Philosophy of Science*, *72*(2), 485–510.
Taylor, C. (1985). *Philosophical papers: Volume 1, Human agency and language*. Cambridge University Press.
Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
Thoma, J. (2021). On the possibility of an anti-paternalist behavioural welfare economics. *Journal of Economic Methodology*, *28*(4), 350–363.
Vetter, B. (2015). *Potentiality: From dispositions to modality*. Oxford University Press.
Wallace, R. J. (2013). *The View from here: On affirmation, attachment, and the limits of regret*. Oxford University Press.
Watson, G. (1975). Free agency. *The Journal of Philosophy*, *72*(8), 205–220.
Weitzman, M. L. (2007). A review of the Stern Review on the economics of climate change. *Journal of Economic Literature*, *45*(3), 703–724.
Williams, B. (1973). A critique of utilitarianism. In J. J. C. Smart, & B. Williams (Eds.), *Utilitarianism: For and against*. Cambridge University Press.