Is Consciousness Intrinsically Valuable?

ANDREW Y. LEE

NEW YORK UNIVERSITY

ABSTRACT:

There are some things that we think are *intrinsically valuable*, or valuable for their own sake. Is consciousness——subjective, qualitative experience—one of those things? Some theorists favor the *positive view*, according to which consciousness is intrinsically valuable. According to a positive theorist, consciousness itself accrues intrinsic value, independent of the particular kind of experience instantiated. In contrast, I favor the *neutral view*, according to which consciousness is neither intrinsically valuable nor disvaluable. The primary purpose of this paper is to clarify what is at stake when we ask whether consciousness is intrinsically valuable, to carve out the theoretical space, and to evaluate the question rigorously. The secondary purpose is to show why the neutral view is attractive and why certain arguments for the positive view do not work.

---

§0 | IS CONSCIOUSNESS INTRINSICALLY VALUABLE?[1]

There are some things that we think are valuable, though not for their own sake. Money might be valuable in this way—money seems valuable only in virtue of the fact that it can be used to purchase goods. Things that are valuable in this way are *instrumentally valuable*. There are other things that we think are valuable for their own sake. Pleasure might be valuable in this way—there seems nothing further in virtue of which pleasure is valuable. Things that are valuable in this way are *intrinsically valuable*. Is consciousness—subjective, qualitative experience—itself intrinsically valuable?

Many theorists say yes. For example, Frankena [1973] includes consciousness in his list of intrinsic goods, alongside venerated goods such as happiness, knowledge, friendship, and beauty. Siewert [1998] argues that with some reflection we can clearly see that "we value consciousness for its own sake." Levy [2014] says "It is widely assumed that consciousness *matters*...If [a] fetus is conscious, it is widely held, its moral value is greater than if it is not yet conscious". Glannon [2016] claims that "[m]any of us believe that consciousness is intrinsically valuable."

Let's call the view that consciousness is intrinsically valuable the *positive view*. For a positive theorist, every experience instantiates some intrinsic value in virtue of being conscious. Of course, the intrinsic value of consciousness might be outweighed by the intrinsic disvalue of other factors. But the positive theorist thinks that some intrinsic value is accrued in virtue of consciousness itself, independent of the particular character of the experience. The following passage from Nagel [1979] illustrates the positive view:

> "There are elements which, if added to one's experience, make life better; there are other elements which if added to one's experience, make life worse. But

[1] Many thanks to Matthew Liao, who has provided feedback across multiple drafts of this article. Thanks also for comments from Kyle Blumberg, David Chalmers, Ben Holguin, Rob Hopkins, Arden Koehler, Sam Lee, Adam Lovett, Rob Long, Thomas Nagel, Sam Scheffler, Sharon Street, David Velleman, Jake Zuehl, and audiences at NYU and Institut Jean-Nicod.

> what remains when these are set aside is not merely neutral: it is emphatically positive. Therefore life is worth living even when the bad elements of experience are plentiful, and the good ones too meager to outweigh the bad ones on their own. The additional positive weight is supplied by experience itself, rather than by any of its consequences."
>
> —Thomas Nagel, *Mortal Questions* [1979]

Other theorists are skeptical that consciousness itself is intrinsically valuable. Perhaps capacities often associated with consciousness, such as cognition or sentience, are intrinsically valuable, even if consciousness itself is not. Or perhaps consciousness is merely necessary for certain intrinsically valuable properties, without itself being an intrinsically valuable property.

Let's call the view that consciousness is neither intrinsically valuable nor intrinsically disvaluable the *neutral view*. The neutral theorist might think that particular kinds of experiences are intrinsically valuable, but that these experiences are valuable in virtue of their particular phenomenal character, and not in virtue of their being conscious. The following passage from Glover [2006] expresses sentiments favoring the neutral view:

> "It would be possible to hold mere [phenomenal] consciousness to be of intrinsic value...But when the principle is distinguished from different ones that would place a value on higher levels of consciousness, it has so little intuitive appeal that we may suspect its attractiveness to depend on the distinction [between phenomenal consciousness and more sophisticated forms of consciousness] not being made."
>
> —Jonathan Glover, "The Sanctity of Life" [2006]

I favor the neutral view. But the primary purpose of this paper is to clarify what is at stake when we ask whether consciousness is intrinsically valuable, to carve out the theoretical space, and to evaluate the question rigorously. The secondary purpose is to show why certain arguments for the positive view do not work and why the neutral view is attractive.

§1 | FRAMEWORK

The sense of "consciousness" I am concerned with is *phenomenal consciousness*. A creature is phenomenally conscious just in case there is something it is like to be that creature, or just in case that creature has a first-personal perspective. The *phenomenal properties* of an experience constitute the subjective, qualitative character of that experience, or what it is like to have that experience. For example, the visual experience of red, the feeling of pain, and the flavor experience of umami are all phenomenal properties.

Sometimes consciousness is construed as a property of mental states (such as perceptual states), and sometimes it is construed as a property of creatures (such as humans). A mental state is *state conscious* just in case there is something it is like to be in that mental state, or just in case that mental state has a phenomenal character. A creature is *creature conscious* just in case it has conscious mental states.[2] Both of these notions of consciousness will figure in to the discussion.

Something has *intrinsic value* just in case that thing is good for its own sake. Putting it another way, if something has intrinsic value, then there are no further evaluative grounds to appeal to in explaining why that thing is good.[3] Consider the following characterization of intrinsic value from Zimmerman [2015]:

> "[I]f one thing derives its goodness from some other thing, which derives its goodness from yet a third thing, and so on, there must come a point at which you reach something whose goodness is not derivative in this way, something that "just is" good in its own right, something whose goodness is the source of, and thus

---

[2] Sometimes creature consciousness is instead characterized as the *capacity* to have conscious mental states, but the formulation I use is more relevant for the purposes here.

[3] I'll also assume that if something has intrinsic value, then it necessarily has intrinsic value. Some theorists (e.g., Korsgaard [1983]) disagree with this characterization of intrinsic value, and think that something can have intrinsic value contingently—though these claims are controversial (see Zimmerman [2015] for some discussion). But even if we accept that there could be contingent intrinsic value, we can simply reframe our question as concerned with whether consciousness necessarily has intrinsic value.

> explains, the goodness to be found in all the other things that precede it on the list. It is at this point that you will have arrived at intrinsic [value]"
>
> —Zimmerman, "Intrinsic vs. Extrinsic Value" [2015]

The passage above articulates the core notion of intrinsic value. There are some further issues about intrinsic value that I'll remain neutral on. First, I leave open the relationship between value and morality—I make no assumptions or claims about what is morally permissible or obligatory.[4] Second, I leave open what the grounds of intrinsic value might be. For example, my arguments are consistent with taking intrinsic value to be a primitive property, to be grounded in natural physical properties, or to be grounded in the evaluative attitudes of agents. Third, I am concerned with the kind of value that corresponds to the notion of making an agent or a situation better off. Sometimes philosophers are concerned with other kinds of value, such as one that is fundamentally concerned with respect or with rights, but I leave open what the relationship is between those kinds of value and the kind that I am concerned with here.

Along with the concept of intrinsic value comes the contrasting notion of *intrinsic disvalue*. Whereas intrinsic value concerns the good, intrinsic disvalue concerns the bad. Pleasures plausibly have intrinsic value, while pains plausibly have intrinsic disvalue (though we need not assume that whether an experience is valuable or disvaluable depends solely on how much pleasure and pain it instantiates). An experience that is on the whole neither intrinsically valuable nor intrinsically disvaluable is *evaluatively neutral*. There are further questions we could ask about the relationship between intrinsic value and intrinsic disvalue, but the only assumption we need is that intrinsic value and intrinsic disvalue are

---

[4] Some papers have been explicitly concerned with the moral (as opposed to evaluative) significance of consciousness. For example, Levy & Savulescu [2009] argue that consciousness is sufficient for moral patiency (but not for moral personhood). Other papers, such as Siewert [1998, 2013] and Kahane & Savulescu [2009] seem concerned with both moral and evaluative issues. I am explicitly concerned with issues about what makes things better or worse rather than right or wrong, and I make no claims about what bearing my considerations have for the moral significance of consciousness.

commensurable (at least when concerning experiences). This means that the intrinsic value instantiated by an experience can be weighed against the intrinsic disvalue instantiated by an experience to determine which of a range of scenarios is best.

Most theorists accept that the intrinsic value or disvalue of an experience is determined in part by its *valence*, or the degree to which an experience is pleasant or unpleasant. For example, a pleasurable experience of eating sushi has positive valence, and a painful experience of a stomachache has negative valence. Note that the notion of the value of an experience is conceptually distinct from the notion of the valence of the experience. But I'll assume that positive valence is intrinsically valuable while negative valence is intrinsically disvaluable, and I'll assume that there is a positive linear correlation between the respective properties.[5] I'll use the terms *positive* and *negative* to denote valence, and the terms *good* and *bad* to denote value.

The question of whether consciousness is intrinsically valuable can be framed either in terms of *global value*, which concerns whether things in general are better in virtue of a creature being conscious, or in terms of *individual value*, which concerns whether a particular creature is better in virtue of being conscious. My discussion is formulated in terms of global value. But by and large, the arguments apply analogously to individual value.[6]

I'll take the primary bearers of intrinsic value to be properties. Sometimes I talk about the intrinsic value of objects (such as conscious beings), states of affairs, or property instantiations, but these can be understood in a derivative sense, where the intrinsic values of these latter entities are grounded in the intrinsic

---

[5] No part of this paper crucially relies upon these assumptions, but taking these assumptions for granted will make the discussion smoother.

[6] If you think that only creatures that are conscious can be bearers of individual value, then perhaps we cannot ask whether is creature is better off if conscious because there are no scenarios in which the creature exists but is not conscious. But even if this is right, we could still ask whether intrinsic value is accrued in virtue of consciousness itself, or whether all intrinsic individual value is accrued in virtue of other properties.

value of the properties they instantiate (or are instantiations of). Furthermore, even a theorist who thinks that the fundamental bearers of intrinsic value are objects or states of affairs must still think that there are value-making properties in virtue of which objects or states of affairs have intrinsic value. For these theorists, we can translate the question of whether consciousness is intrinsically valuable to the question of whether consciousness is a value-making property.

Our question is distinct from the question of whether consciousness is sufficient for intrinsic value. Perhaps it is necessarily the case that all conscious beings have some intrinsic value due to some other property that is always instantiated when consciousness is instantiated (perhaps structural complexity or valence), even if consciousness itself is not intrinsically valuable. What we are interested in is whether something is intrinsically valuable in virtue of instantiating consciousness.

Our question is also distinct from the question of whether consciousness is necessary for intrinsic value.[7] There might be some properties that are intrinsically valuable and necessarily entail consciousness (perhaps pleasure), even if consciousness itself is not intrinsically valuable. On a related note, the issue of the intrinsic value of consciousness is orthogonal to the issue of *value experientialism*—the view that all value is grounded in consciousness. Even if value experientialism is true, it might still be that consciousness itself is not intrinsically valuable.

In sum: When asking whether consciousness is intrinsically valuable, we are interested in whether things are better in virtue of consciousness being instantiated. The rest of the paper explores how we might answer this question.

§2 | CONCEPTS

Some theorists have thought that consciousness is intrinsically valuable because it enables many kinds of intrinsically valuable experiences. Without consciousness, there could be no experiences of pleasure, or beauty, or love. But obviously, experiences of these kinds are valuable for their own sake. Even if the same functional and behavioral properties could be preserved in the absence of

[7] See Sidgwick [1907] and Moore [1903] for some classic discussions of this issue.

consciousness, it is still better that that these mental states are conscious. From this, these theorists reach the conclusion that consciousness is intrinsically valuable. The following passage from Seager [2001] (where he is endorsing Siewert [1998]'s arguments that consciousness is intrinsically valuable) illustrates this thought:[8]

> "Consider [a] thought experiment invoked to show that phenomenal consciousness has intrinsic value. The experiment involves imagining a choice between a life with and a life without consciousness...Imagine the devil gives you the choice: you can become the richest and most successful person on the planet, but at the cost of a total loss of consciousness. You will be a zombie, though undetectably such to the rest of the world, and a very well off zombie at that. It is easy to see that, all other things equal, this offer is no bargain; it is tantamount to death."
>
> —Seager, "Consciousness, Value, and Functionalism" [2001]

These theorists seem to be thinking that a property is intrinsically valuable whenever it has some instances that are intrinsically valuable. For example, Siewert [2000] says "I take the question of the intrinsic value of consciousness to be that of whether we value having phenomenal features, but not only for the sake of what other non-phenomenal ones we think will come with them." And Seager [2001] says "The core thought here is that conscious experience is intrinsically valuable, where this means that there are some conscious experiences which are worth having for themselves." But this way of thinking about intrinsic value is untenable, for two reasons.

First, this conception of intrinsic value is too permissive. If all that it takes for a property to be intrinsically valuable is that it has some intrinsically valuable

---

[8] Some might worry about the fact that these considerations require us to consider metaphysical impossibilities. According to many theories of consciousness, we cannot hold fixed all the physical features of a being while varying whether or not that being is conscious. But it is epistemic possibility, rather than metaphysical possibility, that is relevant for normative theorizing. For example, we can (and do) consider the normative upshot of actual lobsters being conscious versus the normative upshot of actual lobsters being non-conscious, even if we think that one of these possibilities must be metaphysically impossible.

determinates, then it becomes nearly trivial that some highly determinable properties are intrinsically valuable. For example, this would likely entail that existence is intrinsically valuable, since it is plausible that there are many ways of existing that are intrinsically valuable (such as being in pleasure). But whether existence is intrinsically valuable is a substantive and unobvious thesis, not a nearly trivial claim. By adopting this liberal notion of intrinsic value, we relinquish one of the major constraints on what it takes for a property to be intrinsically valuable.

The second problem is that this conception of intrinsic value has the uncomfortable result that consciousness is also intrinsically disvaluable. If we accept that any property that has intrinsically valuable instances is itself intrinsically valuable, then it's reasonable to adopt an analogous notion of intrinsic disvalue. Since there are clearly instances of consciousness that are intrinsically disvaluable, such as experiences of extreme suffering, consciousness would then also be an intrinsically disvaluable property. But this is highly counterintuitive, and unlikely to be a consequence that a positive theorist would want to adopt. For these reasons, we should be wary of adopting a notion of intrinsic value that takes any property that has intrinsically valuable instances to itself be intrinsically valuable.

There is a more general lesson to be drawn here. There may be ways of massaging the concept of intrinsic value that render the previous arguments sound.[9] But it's hard to see how such a maneuver could avoid entailing that far more properties are intrinsically valuable (and intrinsically disvaluable) than we might like. On the other hand, if we stick with a notion of intrinsic value that is more

---

[9] A related worry concerns semantics, rather than concepts. Perhaps there is only one concept of intrinsic value, but sentences of the form 'X is intrinsically valuable' can sometimes mean 'Some determinates of X are intrinsically valuable.' But this leads to worries analogous to those just discussed. For example, this view makes the implausible prediction that there are readings of sentences such as 'Existence is intrinsically valuable' that are obviously true. Moreover, if this semantic analysis applies to sentences of the form 'X is intrinsically valuable', then it plausibly also applies to sentences of the form 'X is intrinsically disvaluable.' But this would entail the implausible result that there is a true reading of 'Consciousness is intrinsically disvaluable.'

substantial, the arguments from earlier clearly do not work. In other words, the arguments from before are either trivial or fallacious. Just because some conscious experiences are intrinsically valuable does not mean that consciousness itself is.

To motivate the positive view, the positive theorist needs arguments that are not susceptible to the symmetry considerations discussed in this section. The next section considers some of the options for the positive theorist.

§3 | SYMMETRY

Some forms of consciousness are good, and some forms are bad. The question for the positive theorist is why we should think that consciousness itself is intrinsically valuable in spite of the axiological symmetry between its determinates.

These symmetry considerations are illustrated more vividly when we consider Hell—a world inhabited by beings whose lives are never-ending onslaughts of pain, fear, despair, and failure. When the philosopher in Hell thinks about the axiology of consciousness, they might be drawn towards neither the positive view nor the neutral view, but instead the *negative view*, according to which consciousness is intrinsically disvaluable. After all, zombies have the good fortune of foregoing the many unpleasant experiences that conscious beings must suffer through. Of course, this inference seems unwarranted to us. But then this raises the question of why the positive theorist is justified while the negative theorist is wrong.

It's worth noting that many canonical examples of intrinsic properties, such as knowledge, pleasure, and beauty, do not exhibit these symmetries. For example, although some kinds of knowledge may lead to bad results, it's hard to think of forms of knowledge that are plausibly bad for their own sake. And even those who think there might be some exceptions to the rule (perhaps sadistic pleasure is a bad form of pleasure that is not intrinsically valuable) tend to think that prototypical determinates of the properties under consideration are intrinsically valuable. But in the case of consciousness, negative experiences have just as much claim to being prototypical cases of consciousness as positive experiences.

What moves might the positive theorist make to break the symmetry? One kind of move is to draw an analogy with other candidates for intrinsic goods. For example, consider achievement. Some achievements are good (such as successfully saving a cat from a tree), while other achievements are bad (such as successfully murdering that cat). But some think achievement is intrinsically good, despite being axiologically symmetric. Such theorists think that even achievements of bad ends accrue some intrinsic value in virtue of the achievement itself. Perhaps consciousness behaves similarly.

This might help elicit the positive theorist's intuitions, and this comparison puts consciousness in the company of other candidates for intrinsic goods. However, it's not clear it provides any additional justificatory force. Someone who is uncertain why we should ascribe intrinsic value to consciousness itself when there are both good and bad forms of consciousness is likely to the think that the same applies to achievement. If there are good achievements and bad achievements, and neither have better claim to being canonical cases of achievement, then why should we think that achievement itself lies on the good side? Putting it another way, this may be a case where one person's modus ponens is another's modus tollens.

Another kind of move the positive theorist might make to break the symmetry is to appeal to special properties of consciousness that justify thinking that it's intrinsically valuable. Perhaps consciousness is intrinsically valuable because it enables subjectivity. Without consciousness, there are no points of view—a world without consciousness is a world full of darkness. Or perhaps consciousness is intrinsically valuable because it's a mysterious and marvelous property. It seems almost magical that consciousness could arise from mere physical matter.

But again, while these considerations might draw out the positive theorist's intuitions, it's not obvious that they carry any justificatory force. As an analogy, consider light. Suppose a theorist of antiquity claims that light is intrinsically valuable. After all, light is what enables us to see at all, even if some things we see are ugly or bad. Light is mysterious—it seems radically different other kinds of physical phenomena. Light is marvelous—a ray of light shining down from the sky seems almost magical. And a world without light is—literally—a world full of darkness. By appeal to these facts, this theorist of antiquity tries to elicit the

intuition that light is intrinsically valuable. But this seems clearly mistaken. Even though light enables certain significant properties and even though it might be mysterious and marvelous, there is little reason to think that it's intrinsically valuable. But then the question is why these kinds of factors should have a different justificatory status in the case of consciousness.

In my view, the best way for the positive theorist to justify the symmetry break is by simply appealing to core intuitions. Oftentimes, the best that can be done to adjudicate disputes about intrinsic goods is to ensure that we are thinking carefully about the issue at hand and that we are cleanly isolating the relevant intuitions. The current and previous sections have aimed to do the former. The next section aims to do the latter.

§4 | ISOLATION

To test whether a property is intrinsically valuable, we can consider whether varying that property while keeping other factors fixed makes a difference to our evaluative intuitions. In other words, we can *isolate* that property. This section develops three kinds of thought-experiments designed to isolate consciousness. By examining these thought-experiments, we'll also see how to develop several different versions of the positive view.

NEUTRAL CASE

Since consciousness is a determinable property, we can't completely isolate it. In order for consciousness to be instantiated, it must be instantiated in some particular way. But we can strip away as many particular phenomenal features of an experience as possible, and consider whether what remains is intrinsically valuable. Our first thought-experiment aims to do this.

Consider two worlds that are empty save for a single creature inhabiting each world. In the first world, the creature has a maximally simple conscious experience that lacks any valence. Perhaps, for example, the creature has an experience of slight brightness. The creature's experience is exhausted by this sparse phenomenology. In the second world, the creature is not conscious at all. For example, we might suppose that in the second world, the creature is constantly

in a dreamless sleep for the entire duration of its existence. We can stipulate that the two worlds are as similar as possible without violating the difference in consciousness between the two creatures.

Is either of these worlds better? It's certainly not obvious if so. For those who think that intrinsic value entails pro tanto reasons for action, it's not obvious that it would be better for one to create the first world over the second world.[10] And if we had the choice between creating experiences of the kind experienced by the creature or increasing the pleasure of existing experiences, it's not obvious that we could justify on evaluative grounds the former option over the latter.

It's also worth ensuring that our intuitions are genuinely evaluative intuitions, rather than merely preferential intuitions. Consider again the analogy with light. Suppose there are two worlds, the first of which is full of light and the second of which is completely dark. There are many who might find the first world preferable to the second. But as mentioned in the last section, there are few who would want to claim that light is intrinsically valuable. To track the question of whether consciousness is intrinsically valuable, intuitions about this scenario must be different in kind than those that one might have about light.

Last, it's worth ensuring that one's intuitions are genuinely tracking the evaluative status of consciousness itself, rather than of the particular brightness experience of the creature. To some extent, we can abstract away from this worry by considering other kinds of experiences that seem evaluatively neutral (for example, experiences of a low rumble, or saltiness, or a bare tactile sensation). If intuitions remain robust across these cases, then that is some reason to think that it is consciousness itself that is driving the intuitions.[11]

If you think that the consciousness world is better than the non-consciousness world even after taking these methodological precautions, then you favor the positive view. On the other hand, if you think that there is nothing better

---

[10] If the idea of creating an actual universe is too far-fetched, we could imagine that we are able to create simulated worlds with conscious inhabitants.

[11] Nevertheless, someone who thinks that a sparse brightness phenomenal character is intrinsically valuable might likewise be inclined to think that these other sparse phenomenal characters are intrinsically valuable.

about the world where the creature is conscious, then you favor the neutral view. Speaking for myself, once I disentangle my preferential intuitions and focus on the evaluative issues, I find it hard to see why we should think the world with consciousness is better than the world without.

PAIN CASE

If the positive view is correct, then we might wonder how much intrinsic value consciousness instantiates. For example, the Nagel [1970] passage mentioned at the beginning of this paper claims that the intrinsic value of consciousness is "emphatically positive", suggesting that even a life full of suffering can be overall good.

A positive theorist can test how much intrinsic value consciousness accrues by weighing it against the intrinsic disvalue accrued by a negative phenomenal character. Unless the positive theorist takes consciousness to accrue a trivially small amount of intrinsic value (an option I'll discuss later), there should be some negative phenomenal characters whose intrinsic disvalue outweighs the intrinsic value of consciousness. In contrast, a neutral theorist must think that there are no cases where this condition is satisfied.

Consider again two worlds that are empty save for a single creature inhabiting the universe. In the first world, the creature has a painful experience, and the character of its experience is fully exhausted by this pain phenomenology. We can modulate the magnitude of pain in order to test the intrinsic value of consciousness.[12] In the second world, the creature is not conscious at all. And again, suppose that the two creatures are as similar as possible without violating this difference in consciousness.

By modulating the magnitude of pain felt by the creature in the first world and evaluating when the two worlds are equally good, a positive theorist can determine how much intrinsic value they want to ascribe to consciousness.[13] For

[12] I'll assume that intrinsic disvalue of pain correlates linearly with magnitude of pain.

[13] This could also be done using just a single world with a conscious creature and modulating the creature's level of pain until the world is evaluatively neutral. However, the pairwise comparison is better for isolating the evaluative contribution of consciousness

example, suppose you think that the two worlds are equally good when the creature in the first world feels an intense migraine headache. This would indicate that the intrinsic value of consciousness is equal to the intrinsic disvalue of the intense migraine experience. Of course, this only gives us a comparative value, rather than an absolute value. But this is still useful, given that we can have relatively clear grips on how bad various levels of pain are.

It's worth mentioning that even if you think that the world where the creature is not conscious is better for some magnitudes of pain, you need not think that it is always better for creatures suffering to that degree to not be conscious. Consider a terminally ill person whose suffering is permanent. You might think that it is nevertheless better if that patient is conscious, because there are other features of that person's experiences that are valuable. For example, there may be value in remembering one's past, in interacting with other people, in feeling satisfaction about one's accomplishments, or in having rich or interesting thoughts. But we can stipulate that the creature in the above scenario is not suffering for the sake of anything, is suffering perpetually, and has none of the kinds of experiences we typically think of as valuable. By making these stipulations, we isolate the factors relevant to the disagreement between the positive theorist and the neutral theorist.

Different positive theorists will take different stances on what magnitude of pain makes the two worlds evaluatively equal. Some, such as Nagel, think that the magnitude of pain must be very high to counterbalance the intrinsic value of consciousness. These theorists might even think that Hell is a better world than a similar world without consciousness. Others might think that the magnitude of pain must be low or moderate to counterbalance the intrinsic value of consciousness. These theorists might think that a world with a creature experiencing only a mild pain is better than a maximally similar world with no consciousness at all.

We can call the aforementioned options *substantial positive views*, since they ascribe a non-trivial amount of intrinsic value to consciousness. Many theorists

---

because it abstracts away from other potentially valuable features (such as the creature itself, independent of its experience).

are likely to find substantial positive views quite counterintuitive. Suppose a dental patient must undergo a root canal, and they can be either conscious or unconscious during the operation. Few would want to say that it's better if the patient is conscious. Of course, this is a scenario where the patient is unconscious only temporarily—perhaps there is a relevant difference between cases concerning a single experience amongst many and cases concerning the entire life of a creature. But we could modify the dental surgery scenario to so that it ranges over the entire life of a (quite unfortunate) person. If anything, it seems even harder to see why it would be better that the person experiences dental surgery when those experiences last the person's entire life.

Of course, a substantial positive theorist need not think that consciousness has so much intrinsic value as to outweigh the disvalue from the painful dental surgery experience. Some might think that only milder cases involving milder pains are ones where the good of consciousness is counterbalanced by the bad of the pain. Speaking for myself, I find it counterintuitive that it could ever be better that a creature is in pain than not conscious at all, other things being equal. The severity of the cases differ in degree, but they do not seem to differ in kind. If your intuitions diverge here, then we may have reached the point of intuitional bedrock.

There are a few ways that positive theorists who share my intuitions might respond. One option concerns *incommensurability*. Perhaps the intrinsic value of consciousness is incommensurable with the intrinsic disvalue of the specific phenomenal character. If this is correct, then we cannot claim that either world is better or worse simpliciter. But this response is unmotivated. There are cases where the intrinsic value of consciousness is clearly commensurable with the intrinsic disvalue of a specific phenomenal character. Suppose, for example, that if the creature is conscious, then it perpetually experiences torturous, excruciating pain and massive anxiety. It's implausible that in such a case we simply cannot assess which world is better. This suggests that the intrinsic value of consciousness is commensurable with the intrinsic disvalue of the specific phenomenal character after all.

Another option concerns *undercutting*. Some think that in certain circumstances, the intrinsic value of a property can be undercut, meaning it no longer generates intrinsic value in that circumstance. For example, such theorists

might think that pleasure is intrinsically valuable, but its intrinsic value is undercut if one takes pleasure in the misfortune of others. Perhaps similarly, the intrinsic value of consciousness is undercut when consciousness has a negative phenomenal character. It's controversial whether undercutting is possible at all. But even if we grant that undercutting is possible, the diagnosis seems implausible in the case of negative experiences. Purported examples of undercutting, such as the example of taking pleasure in someone else's misfortunes, are typically taken to be exceptional cases. But it's implausible to think that negative experiences are exceptional cases of consciousness whereas positive experiences are the prototypical cases. Perhaps the positive theorist might argue that undercutting does not require exceptional cases and that it is tenable to hold that all negative experiences involve undercutting. Or they might argue that there is something distinctive about the pain case that renders it a case of undercutting, even if not all cases of negative experiences involve undercutting. Neither of these responses seems promising to me, but the details will depend on the particular account that the positive theorist provides.

Perhaps a more attractive option is the *minimal positive view,* according to which consciousness is intrinsically valuable but where its intrinsic value is trivially small. This theory would predict that a world containing a creature who has experiences with neutral phenomenal character is better than a maximally similar world without consciousness (the scenario of the previous subsection), but that a world with negative phenomenal character would always be worse than a maximally similar world without consciousness (the scenario of the current subsection). The minimal positive theory holds that even the faintest and briefest pain is enough to outweigh the intrinsic value of consciousness.

However, if you are drawn towards the minimal positive view, then I think there is extra reason to be scrupulous about intuitions here. When intuitions are weak or minimal, there is greater risk of conflating preferential intuitions with evaluative intuitions, as discussed previously. On the other hand, if your intuitions about which world is better are clearly tracking evaluative factors, then the minimal positive view is a way of retaining the claim that consciousness is intrinsically valuable while avoiding the counterintuitive consequences of the substantive positive views. Speaking for myself, I think the minimal positive view

is more attractive than the substantive positive view. But it also takes the intrinsic value of consciousness to be negligible, rendering the issue less significant than one might have thought.

Where does this leave the dialectic? Substantial positive views lead (at least in my mind) to quite implausible consequences. Minimal positive views are more plausible, but they render the intrinsic value of consciousness trivially small. But in what follows, I'll discuss a third version of the positive view that avoids the counterintuitive consequences of the substantial views while retaining the substantiality of intrinsic value of consciousness.

QUANTITY CASE

If consciousness is intrinsically valuable, does every conscious experience generate the same amount of intrinsic value in virtue of consciousness itself? So far, we've assumed that the intrinsic value of consciousness is constant. But perhaps this assumption oversimplifies. Some positive theorists might think that the more conscious a creature is, the more intrinsic value its experience accrues. Just as the value of a pile of gold is a function of the quantity of gold, perhaps too the value of a conscious experience is a function of the quantity of consciousness. A *scalar positive theorist* takes the intrinsic value of consciousness to scale with how conscious a creature is.

By adopting a scalar positive view, the positive theorist can avoid both the counterintuitive consequences of the substantial positive views and the triviality of the minimal positive view. If there is a low amount of consciousness—perhaps in the case of a fly—there is little intrinsic value instantiated. If there is a high amount of consciousness—perhaps in the case of a human—then there is much more intrinsic value instantiated. Whereas the value of the former experience might be outweighed by the disvalue of even a slight pain, the value of the latter experience might be outweighed only by the disvalue of extremely intense pains. But all this raises a question: can we develop an account of how consciousness scales?

It's certainly not obvious how to scale consciousness. In the case of gold, there are quantifiable units that we can use for measuring how much gold there is. But we have no clear candidates for what those units might be for consciousness. Unless we have a principled way of making sense of quantity of consciousness, the

scalar view cannot get off the ground.[14] Moreover, many theorists are skeptical of the notion of quantity of consciousness. Bayne, Hohwy, & Owen [2016], for example, express skepticism towards the idea that there is a single, principled way of quantifying the degree of consciousness. Kahane & Savulescu [2009] express similar thoughts, claiming that "it is far from clear that we can coherently speak of phenomenal consciousness as a matter of degree." So, there is a substantive question of whether the scalar positive theorist could develop a viable account of the quantity of consciousness.[15]

I can see two different potential approaches to such an account. One possibility requires that we adopt an atomist view of the structure of experience, where total experiences are composed from more basic atomic experiences. If atomism is correct, then we might understand quantity of consciousness as a function of the number of atomic experiences that compose a total experience. Another possibility is to understand quantity of consciousness in terms of how many degrees of freedom an experience has. This approach might analyze quantity of consciousness as a function of how many different dimensions of variation a creature's experience instantiates.

Thinking in detail about how an account of quantity of consciousness could be developed would take us too far astray from the main path. And even if we did

---

[14] Another way to quantify consciousness is to simply aggregate subjects of experience. There is more consciousness in a world that contains ten subjects than a world that contains just one subject. It's relatively easy to move from the single subject cases considered previously to multi-subject cases, and I suspect that most people's intuitions will remain stable as we generalize to the multi-subject cases. On the other hand, thinking about quantity of consciousness within a subject raises new and interesting theoretical considerations. For these reasons, I focus on the intrasubject rather than intersubject cases here.

[15] One prominent account of quantity of consciousness comes from integrated information theory (Tononi [2008]). But it is unclear how exactly to interpret what quantity of consciousness is measuring under integrated information theory (see Pautz [unpublished] for a critical analysis). Examining how integrated information theory interacts with our question about the intrinsic value of consciousness would be interesting, but would also take us too far from the core issues of this paper.

have an account of what quantity of consciousness consists in, that still leaves open what kinds of creatures actually instantiate high quantities of consciousness. Instead, suppose we grant the scalar positive theorist the assumption that there is a viable account of quantity of consciousness that might be taken to scale with the intrinsic value of consciousness.

If we take consciousness to scale in quantity, then there are further thought-experiments we can develop to think about the intrinsic value of consciousness. Whereas our first-thought experiment involved zero degrees of freedom and our second thought-experiment involved one degree of freedom (namely, the intrinsic disvalue from the pain experience), our third thought-experiment involves two degrees of freedom. The first is the intrinsic disvalue of pain (modulated by the magnitude of pain). The second is the intrinsic value of consciousness (modulated by the quantity of consciousness). By examining how these parameters interact, a scalar positive theorist can test how the intrinsic value of consciousness scales.

Consider again two worlds. The first world contains a conscious creature that has a pain experience. The first parameter we can vary is how conscious that creature is. The second parameter is the magnitude of that creature's pain. We can stipulate that outside of the creature's pain experience, the rest of the phenomenal character of the creature's experience is on balance evaluatively neutral. By characterizing the first world in this way, we can leave the phenomenal character of the creature's experience open enough to accommodate a variety of views about how to quantify the amount of consciousness while still isolating the variables that we're interested in examining. In the second world, the creature is not conscious at all. Once again, suppose that the two creatures are as similar as possible without violating this difference in consciousness.

Since we don't have a specific account of what quantity of consciousness consists in, we are not in a position to make clear judgments in case. But this methodology gives the positive theorist a way of gauging how the intrinsic value of consciousness scales. The positive theorist can examine which magnitudes of pain and which quantities of consciousness are such that the world with the conscious creature is equally good to the world with the non-conscious creature. By doing so, they can determine which quantities of consciousness evaluatively cancel out which magnitudes of pain.

There are further general constraints we can think about for the scalar positive view, independent of any particular implementations. For example, there is the question of whether there could be an experience such that the intrinsic disvalue of its specific phenomenal character would always outweigh the intrinsic value of consciousness. Is there an experience so bad that it would be better if the creature were non-conscious, no matter how conscious the creature is?[16] If the scalar positive theorist says no, then they must hold that in some cases it is better if a creature suffers terribly even though there are no countervailing positive experiences. On the other hand, if the scalar positive theorist says yes, then they would think that there is a ceiling on how much intrinsic value could be accrued from consciousness. If this latter view is correct, the function from quantity of consciousness to intrinsic value is asymptotic.

I myself find the scalar positive view to be the most attractive version of the positive view. It avoids the counterintuitive consequences of the substantive positive views while avoiding the triviality of the minimal positive view. But the scalar positive view is viable only if we have a viable account of quantity of consciousness. This is a real challenge for the scalar view, especially since there is no current consensus on how to quantify consciousness, or whether such a notion is defensible in the first place. And this also leads to a surprising result—the best way to hold that consciousness is intrinsically valuable requires endorsing some substantive claims about the structure of experience.

§5 | CLOSING REMARKS

Some, upon engaging with the contents of this paper, have thought that the neutral view is obviously correct. Others have thought that the neutral view is defensible, but that the positive view is more intuitive. And others have found themselves unsure of what to think in the end. Through conversations and comments, I have encountered people in each of those categories. It may seem obvious that there are readers of all stripes. But I mention these remarks in order to address two kinds of criticisms that this paper occasionally receives. The first

[16] One might also think that there are limits on how conscious a creature could be, depending on the details of one's account of amount of consciousness.

criticism is that almost everyone who has thought about the issue clearly would endorse the neutral view, and that the positive view is a strawman. The second criticism is that I have not sufficiently motivated the neutral view and that the intuitions favoring the positive view are more compelling than what I suggest.

Of course, there is a tension between the two criticisms. Taken in conjunction, they suggest that in fact the issue is not so obvious either way. When there is disagreement at the level of bedrock intuitions, there may be little room for further analysis beyond ensuring that our intuitions do not stem from a dubious source. In light of this, I've aimed to strike a balance between voicing my own views on the issue and articulating the best versions of the positive view. Even when there is a clash of fundamental intuitions, we can at least ensure that both sides are clear on what kind of view they are committed to and why. Based off of my own experience, it is clear that there are people on both sides of the divide.

Whether or not consciousness is intrinsically valuable is an interesting question in its own right. But it might also have implications for other issues in value theory. Consider the question of what the threshold is for a life worth living. The Nagel passage mentioned at the beginning expressed the sentiment that life is worth living even when the bad experiences outweigh the good experiences, because there is additional value provided by consciousness itself. If the neutral view is correct, however, then consciousness does not confer any additional value, and a life where the bad otherwise outweighs the good would not be worth living. Or consider the question of whether being conscious entails having moral status. If the neutral view is correct, and if we think that having moral status requires instantiating at least some evaluatively significant properties, then merely being conscious might not be enough to make a being worthy of moral consideration. Or suppose that in the future we learn how to create conscious artificial intelligence. Then issues about the intrinsic value of consciousness may very well have practical implications for how we ought to act with respect to such beings.

Though this paper has focused on consciousness, the methodology could also be applied to other determinable properties that are thought to be intrinsically valuable. When considering whether some property is intrinsically valuable, it's important to avoid confusing the intrinsic value accrued in virtue of the instantiation of that property itself with the intrinsic value accrued in virtue of the

specific way in which that property is instantiated. Only if our intuitions remain stable after considering the property in a wide range of its instantiations should we think that the property is intrinsically valuable. Other examples of determinable properties that are sometimes taken to be intrinsically valuable and where analogous arguments might hold include life and existence. Just as in the case of consciousness, we should consider a wide range of ways that things can be alive and ways that things can exist, and not focus merely on the good cases.

Although my main focus throughout the paper has been to develop different versions of the positive view and to evaluate different arguments for it, I hope that the attraction of the neutral view has also become manifest. According to my preferred version of the neutral view, consciousness is an enabler of intrinsic value but is not itself intrinsically valuable. Rather, consciousness is a determinable property that has both valuable and disvaluable determinates. Under this view, there is symmetry between positive and negative experiences—positive experiences are valuable because of their specific phenomenal characters, and negative experiences are disvaluable because of their specific phenomenal characters. Particular kinds of experiences may be intrinsically good or bad, but consciousness itself is neutral.

REFERENCES

Bayne, Tim ; Hohwy, Jakob & Owen, Adrian M. (2016). Are There Levels of Consciousness? Trends in Cognitive Sciences 20 (6):405-413.

Benatar, David (2006). Better Never to Have Been: The Harm of Coming Into Existence. New York; Oxford University Press.

Crisp, Roger, "Well-Being", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/well-being/>.

Dennett, Daniel C. (1991). Consciousness Explained. Penguin.

Frankena, William K., 1973, *Ethics*, second edition, Englewood Cliffs: Prentice Hall. [1, 6]

Glannon, Walter (2016). The Value and Disvalue of Consciousness. *Cambridge Quarterly of Healthcare Ethics* 25 (4):600-612.

Gligorov, Nada (2008). Unconscious pain. *American Journal of Bioethics* 8 (9):27 – 28.

Glover, Jonathan (2006). The sanctity of life. In Helga Kuhse & Peter Singer (eds.), *Bioethics: An Anthology*. Blackwell. pp. 266--275.

Hsieh, Nien-hê, "Incommensurable Values", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/value-incommensurable/>.

Jaworska, Agnieszka and Tannenbaum, Julie, "The Grounds of Moral Status", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2013/entries/grounds-moral-status/>.

Kahane, Guy & Savulescu, Julian (2009). Brain-Damaged Patients and the Moral Significance of Consciousness. *Journal of Medicine and Philosophy* 34 (1):6-26.

Korsgaard, Christine, 1983, "Two Distinctions in Goodness", Philosophical Review, 92: 169–95.

Levy, Neil & Savulescu, Julian (2009). Moral significance of phenomenal consciousness. Progress in Brain Research.

Levy, Neil (2014). The Value of Consciousness. *Journal of Consciousness Studies* 21 (1-2):127-138.

Moore, G. E. (1903). *Principia Ethica*. Dover Publications.

Nagel, Thomas (1970). Death. Nous. 4 (1):73-80.

Nagel, Thomas (1979). *Mortal Questions*. Cambridge University Press.

Pautz, Adam, What is Integrated Information Theory a Theory Of? (unpublished manuscript)

Seager, William E. (2001). Consciousness, value and functionalism. Psyche 7 (20).

Schroeder, Mark, "Value Theory", The Stanford Encyclopedia of Philosophy (Summer 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2012/entries/value-theory/>.

Siewert, Charles (1998). The Significance of Consciousness. Princeton University Press.

Siewert, Charles (2000). Precis of *The Significance of Consciousness*. *Psyche* 6 (12).

Siewert, Charles (2013). Speaking Up For Consciousness. In Kriegel (ed.), *Current Controversies in Philosophy of Mind*. Routledge.

Sidgwick, Henry, 1907, *The Methods of Ethics*, London, Macmillan, 7th edition.

Tononi, Giulio (2008). Consciousness as Integrated Information: A Provisional Manifesto. The Biological Bulletin 215 (3).

Zimmerman, Michael J., "Intrinsic vs. Extrinsic Value", *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/spr2015/entries/value-intrinsic-extrinsic/>.