*Note: this is the penultimate draft; please cite from the published version.

A Defense of Modest Ideal Observer Theory: The Case of Adam Smith's Impartial Spectator

Nir Ben-Moshe
University of Illinois at Urbana-Champaign

Abstract: I build on Adam Smith's account of the impartial spectator in *The Theory of Moral Sentiments* in order to offer a modest ideal observer theory of moral judgment that is adequate in the following sense: the account specifies the hypothetical conditions that guarantee the authoritativeness of an agent's (or agents') responses in constituting the standard in question, and, if an actual agent or an actual community of agents are not under those conditions, their responses are not authoritative in setting this standard. However, in the account that I provide, the hypothetical conditions can themselves be constructed from the psychology and interactions of actual human beings. In other words, facts about the morally appropriate and inappropriate are determined from hypothetical conditions that—while agents in a given society might have yet to attain them—can be constructed from those agents' shared experiences. Thus, the account offers both an *attainable* standard of moral judgment and a standard that can *transcend* the biases of the society which gave rise to it. I also defend the account against three challenges: (a) ideal observer theories do not offer the right kind of motivation to act on the verdicts of the ideal observer; (b) ideal observer theories cannot explain why the idealization in question is well-motivated and not objectionably ad hoc; (c) the standard used in ideal observer theories cannot be defended upon further reflection, because we would need a non-arbitrary, second-order standard to govern our reflection on the first-order standard of moral judgment.

Keywords: Adam Smith, Anti-Realism, Ideal Observer, Impartial Spectator, Naturalism, Response-Dependence

1. Introduction

For those who embrace a naturalistic picture of the world and are skeptical about the prospects of meta-ethical realism, an answer to the question what accounts for the correctness of moral judgment has proven to be elusive. Assuming one wishes to both (a) ground moral judgment in the *responses* of agents without paying the normative prices of subjectivism and relativism, and (b) attain a robust form of *objectivity* without paying the metaphysical, epistemological, and motivational prices of realism, one could opt for the position that it is the responses of agents *who are under certain conditions* that constitute what is morally right. Since actual agents may not meet these conditions, their responses, individually and collectively, might be in error.[1] Of course, the conditions in question cannot be defined in terms of these agents getting the right results, on pain of circularity. However, these conditions are required to guarantee that the agents' responses are in fact authoritative, and one strategy is to *idealize* them via the postulation of an ideal observer whose reactions determine whether an ethical judgment is true or false.[2] How ideal should the

[1] These observations are based in part on Ben-Moshe (2020b, 431-32) & Vallentyne (1996, 101). For further discussion of this type of non-normative response-dependence, see Lewis (1989) & Johnston (1989). Two points are worth noting. First, there is a normative version of response-dependence theory, according to which it is the *warranted* responses of agents who are under certain conditions that constitute what is morally (or evaluatively) right. See, in particular, McDowell (1998) & Wiggins (1998). I will not be discussing normative response-dependence theories. Second, while I will be arguing in section 5 that the constitutive role of the idealization in Smith's account is well-motivated, I will generally *not* be arguing *against* the view that the responses of agents who are under certain conditions are *evidence of*, and do not *constitute*, what is morally right. See Enoch (2009, 322-23) for a discussion of this distinction in the context of constructivism. Thus, the main argument of this paper is not intended to conclusively rule out the realist possibility, according to which the responses of agents under suitable conditions track stance-independent normative facts. Rather, I am appealing, to some extent, to those naturalist anti-realists who already find the constitutivist alternative appealing. I am grateful to an anonymous referee for urging me to clarify this point.

[2] This position was summed up by Firth (1952, 321), who argued that "X is P [an ethical predicate]" means "Any ideal observer would react to x in such and such a way under such and such conditions." While I will be using Firth as a foil, since he is most identified with ideal observer theory, I will *not* be following him in using the notion of an ideal observer primarily in order to analyze the *meaning* of moral terms. Rather, my ambitions are closer to those of Brandt (1959, ch. 10), who uses the ideal observer—or the "Qualified Attitude Method," as he calls it—primarily in order to provide a *standard of correctness* of moral judgments. See Carson (1984, 50) for a discussion of this difference.

observer be in order to perform this task? Firth (1952, 333) famously argued that the observer should be, amongst other things, omnipercipient and omniscient with respect to the non-ethical facts, and that "any plausible description of an ideal observer will be a partial description of God." The problem with Firth's suggestion is that it *over*-idealizes the observer and detaches him from our human sensibilities. It thus becomes unclear what epistemic access we have to the reactions of such an observer, or why his decisions should bind us or motivate us.[3] However, the roots of ideal observer theory were different: sentimentalists in the 18th century, especially Hume and Smith, used the idea of spectators' responses under certain conditions to articulate a standard of correctness for moral judgments which, while transcending individual points of view, does not transcend the point of view of human beings. In particular, it is the fact that spectators are under these conditions that makes the objects of their sentiments of approval and disapproval—which arise from sympathetic reactions with the actor or with those affected by his actions[4]—*merit* that approval or disapproval. Accordingly, this type of theory can be dubbed "modest ideal observer theory."

My aim in this paper is to build on Adam Smith's account of the impartial spectator in *The Theory of Moral Sentiments* (hereafter "TMS") in order to offer a modest ideal observer theory of moral judgment that is adequate in the following sense: the account specifies the hypothetical conditions that guarantee the authoritativeness of an agent's (or agents') responses in constituting

[3] See Sayre-McCord (1994, 218) for an articulation of epistemic worries that are commonly raised against ideal observer theories. See also Brandt (1955, 409-10). It is worth noting that Kawall (2006) does try to deal with some of these worries. I am skeptical of some of Kawall's proposed solutions—and of their upshot for the normative import and motivational efficacy of the ideal observer—but engaging with his suggestions is beyond the scope of this paper.
[4] In referring to "sympathy" throughout this paper, I follow Smith's (and Hume's) terminology. We would call the phenomenon "empathy," a term that is a translation of the German *Einfühlung* and was coined only in1909.

the standard in question, and, if an actual agent or an actual community of agents are not under those conditions, their responses are not authoritative in setting this standard. However, in the account that I provide, the hypothetical conditions can themselves be constructed from the psychology and interactions of actual human beings. In other words, facts about the morally appropriate and inappropriate are determined from hypothetical conditions that—while agents in a given society might have yet to attain them—can be constructed from those agents' shared experiences.[5] Thus, the account offers both an *attainable* standard of moral judgment and a standard that can *transcend* the biases of the society which gave rise to it. I proceed as follows. I first provide an overview of Smith's impartial spectator account, but also further develop important aspects of the account, especially the transition from a *societal* impartial spectator to a *universal* impartial spectator (section 2). I then make the case that the impartial spectator account can be construed as a (modest) ideal observer theory (section 3). Finally, I defend the account against three challenges: (a) the motivational challenge—ideal observer theories do not offer the right kind of motivation to act on the verdicts of the ideal observer (section 4); (b) the ad-hocness challenge—ideal observer theories cannot explain why the idealization in question is well-motivated and not objectionably ad hoc (section 5); (c) the standards-for-standards challenge—the standard used in

[5] I use "constructed" here in a second-order sense, that is, in relation to the *conditions* from which the normative judgments in question are made, not in relation to the *normative judgments themselves*. In other words, it is not the case that the normative judgments in question are constructed under suitable conditions—these judgments are correct if agents' responses are made under the suitable conditions—but rather the suitable conditions themselves can be constructed from within a given society. Now, realists have argued that if there is unconstructed normative material at the core of one's view, we have a form of realism in disguise (Enoch (2009, 332) and Shafer-Landau (2003, 42)). Therefore, even if I am using construction in a second-order sense, one could argue that the standpoint that I claim is constructed could have been there all along, but nobody acted on or thought about it. My reply is similar to the one that I noted in fn. 1: my aim is not to conclusively rule out this realist alternative. I am grateful to an anonymous referee for encouraging me to clarify this point. Of course, there might be a worry that the conditions themselves sneak in certain normative dimensions that would make the entire account viciously circular. I discuss this worry in section 3.

ideal observer theories cannot be defended upon further reflection, because we would need a non-arbitrary, second-order standard to govern our reflection on the first-order standard of moral judgment (section 6).

2. A Smithian Account of the Impartial Spectator

Smith's sentimentalist story starts with our patterns of approval and disapproval of other agents' actions, given our sympathetic reactions to those agents or those affected by them. According to Smith's model of approbation and disapprobation, when we recognize that there is concordance between our sympathetic passion and the original passion of the agent in question, a sentiment of approval arises; and when we recognize that there is a lack of concordance in passions, a sentiment of disapproval arises (TMS I.i.3.1, I.iii.1.9, & II.i.5.11).[6] However, it would be a mistake to equate *moral judgment* with *approbation*, for if the latter is a psychological response then it seems to be subjective in ways in which we would hope the former is not: when we talk about the correctness of moral judgment, we are interested not only in the question whether we approve [disapprove] of X, but also in the question whether that X *merits* our approval [disapproval]. Smith uses the impartial spectator as a privileged standpoint from which the "passions of human nature" become "proper" (TMS II.i.2.2). Importantly, this standpoint, from which we judge both others *and* ourselves—Smith calls the impartial spectator "conscience" (TMS III.3.4, III.3.29, & III.4.4)—can itself be constructed in a given society. In particular, Smith observed that it is part and parcel of human life that we judge others and find others judging us, that is, that people in human society

[6] TMS is referenced with the relevant part, section, chapter, and paragraph in the Glasgow Edition (Smith 1976).

mirror each other. This allows us to see ourselves through the eyes of others, by internalizing the way in which others respond to us, and thus to make judgments of propriety and impropriety of our own sentiments (TMS III.1.3-5). However, agents in a society might come to realize that the actual spectators who judge them are biased, either because they are not informed about the relevant facts or because they have a personal stake in the circumstances, and are thus unreliable sources for determining what is worthy of approval (TMS III.2.4-5). This realization is a product of our desire to be *worthy* of approval: we are the type of creature that does not merely desire praise and dread blame, but that comes to desire being praise-*worthy* and dread being blame-*worthy* (TMS III.2.1). Hence, Smith argues that agents will seek to go beyond the actual bystanders they encounter and use their *imagination* to create an impartial spectator: "We endeavour to examine our own conduct as we imagine any other fair and impartial spectator would examine it" (TMS III.1.2). More specifically, the impartial spectator is "gradually formed from [our] observations upon the character and conduct both of [ourselves] and of other people" (TMS VI.iii.25): we use our imagination to build on our interactions with others and construct an image of a well-informed and impartial bystander.

Some additional features of the development of the impartial spectator and its achievements are worth noting. The creation of an imagined impartial spectator does not happen *ex nihilo*. First, when we sympathize with others, according to Smith, we imagine being in the situation we take the actor to be in (TMS I.i.1.2 & I.i.1.10-13), which allows us to develop our imaginative capacities. Second, when we repeatedly adopt the points of view of others regarding our conduct, we tend to become more impartial, since our passions as reflected by others are less forceful than our original passions (TMS I.i.4.8). In terms of the achievements of the impartial spectator,

adopting the standpoint of the impartial spectator allows the spectator to humble his self-love, since he sees that he is "but one of the multitude, in no respect better than any other in it" (TMS II.ii.2.1 & III.3.4), and thus allows him to correct his perception of his own interests (which are tied to his self-love) versus the interests of others. The key idea here is simple: if we want to weigh our interests versus someone else's interests, "[w]e must view them, neither from our own place nor yet from his, neither with our own eyes nor yet with his, but from the place and with the eyes of a third person, who has no particular connexion with either, and who judges with impartiality between us" (TMS III.3.3). In other words, we consider other people's interests from a fully informed and impartial point of view. Furthermore, by consulting the impartial spectator, "we […] see what relates to ourselves in its proper shape and dimensions" (TMS III.3.1). That is, the standpoint of the impartial spectator allows the spectator to make a comparison between his and others' interests by affording him an understanding of the aspects of the situation that pertain to himself, but perhaps not to others.[7] By doing so, as well as by humbling our self-love, the impartial spectator allows us to see the situation not only from our own perspective, but, as Smith argues in part VII of TMS, also from other people's perspectives.[8] That is, we are able to imagine not only being ourselves in B's situation, but also being B in B's situation.[9] Therefore, we also see other

[7] As I argue elsewhere, the impartial spectator framework is not supposed to be one in which the spectator is a utility-maximizing device (see, for example, TMS III.3.6). In particular, it does not organize "the desires of all persons into one coherent system of desire," as Rawls (1999, 24) put it, and does not fuse different persons into one person. Rather, when the standpoint of the impartial spectator makes us "see what relates to ourselves in its proper shape and dimensions," it allows us to understand our interests in the context of our own perspectives and compare them to the interests of others in the context of their perspectives, thus respecting the perspectives of individuals (Ben-Moshe (2021)).

[8] Smith believed that our excessive self-love makes it difficult for us to see things from other people's perspectives (TMS III.4.3).

[9] While Smith argues, in part I of TMS, that when we sympathize with B, we imagine how *we* would feel in B's situation (TMS I.i.1.10-13), he argues, in part VII of TMS, that, when we sympathize with B, we imagine how *B*

people's perspectives from a fully informed and impartial point of view. Accordingly, adopting the standpoint of the impartial spectator makes us appreciate that our own interests and perspectives are no more privileged than other people's interests and perspectives, enabling us to take into account the interests and perspectives of all concerned. Specified this way, the standard of the impartial spectator is the type of standard that we associate with moral judgment.

Now, it is not easy to ascertain from TMS whether each culture or society has its own impartial spectator, where members of that group, if they continued to become more impartial and informed, would zero in on *that* spectator, or whether there is a single impartial spectator that emerges from all cultures and societies. On the one hand, Smith was sensitive to the fact that some virtues may be culture-relative: "The different situations of different ages and countries are apt […] to give different characters to the generality of those who live in them, and their sentiments concerning the particular degree of each quality that is either blameable or praiseworthy, vary according to that degree which is usual in their own country and in their own times"; accordingly, the "degree of politeness which would be highly esteemed […] in Russia, would be regarded as rudeness and barbarism at the court of France" and the "degree of order and frugality which, in a Polish nobleman, would be considered as excessive parsimony, would be regarded as extravagance in a citizen of Amsterdam" (TMS V.2.7). Importantly, Smith understands the point about the culture-relativity of virtues not as a mere *description* of states of affairs, but as a *normative* judgment,

would feel in B's situation (TMS VII.iii.1.4). As I demonstrate elsewhere, the *full* development of the latter type of sympathy requires the attainment of the standpoint of the impartial spectator, and so Smith discusses this type of sympathy towards the end of TMS, after he has presented his account of the impartial spectator (Ben-Moshe (2020c)). See also Darwall (1998, 268) for an excellent discussion of the importance of the latter type of sympathy in Smith's moral theory.

noting that "in general the style of manners which takes place in any nation may commonly, upon the whole, be said to be that which is most *suitable* to its situation" (TMS V.2.13; emphasis added). These observations suggest that each culture or society has its own impartial spectator. On the other hand, Smith argues that the impartial spectator is a judge whom we set "between ourselves and those we live with," a person "quite candid and equitable […] who has no particular relation either to ourselves, or to those whose interests are affected by our conduct, […] but is merely a man in general […] the representative of mankind."[10] The latter two phrases suggest that there is one impartial spectator that emerges from all societies and cultures. I think that the strength of Smith's (or a Smithian) account is that it suggests a two-stage approach: First, a *societal* impartial spectator forms when spectators correct misinformation and restrain their self-interest in a way that does not necessarily transcend the biases of their society/culture. Second a *universal* impartial spectator forms, which can now transcend the biases of the society/culture from which it arose.[11]

But now we arrive at a crucial question: how do we get from the idea of an impartial spectator *in* a given society, who applies the society's standards fairly and in knowledge of the facts, to someone who is an impartial spectator *of* the society itself and its standards, and who can judge the social standards themselves? The answer to this question has both a normative and an explanatory dimension. Focusing first on the normative dimension, I wish to suggest that the verdicts of the universal impartial spectator act as *normative constraints* on the specific virtues of a given society or culture, that is, on the verdicts of societal impartial spectators. Indeed, Smith

[10] This quote is taken from a passage which first appeared in the 2nd edition of TMS, remained with minor variations in editions 3-5, and was replaced by a slightly different passage in the 6th edition (TMS III.2.31-32). The quoted passage can be found in a footnote on pp. 129-130 of the Glasgow Edition (Smith 1976).

[11] I am grateful to an anonymous referee for urging me to clarify this point.

himself notes that when it comes to "the general style of conduct or behavior," local virtues cannot authorize substantial "departure from what is the natural propriety of action" (TMS V.2.14). The thought is that societal and cultural norms would be constrained by more general norms prohibiting, for example, slavery and genocide.[12] How does the universal impartial spectator prohibit slavery and genocide? By not arbitrarily limiting the scope of the following claim, which was noted earlier: adopting the standpoint of the impartial spectator makes us appreciate that our own interests and perspectives are no more privileged than other people's interests and perspectives, enabling us to take into account the interests and perspectives of all concerned. In other words, if we focus on a societal impartial spectator, the terms "other people" and "all concerned" might refer only to, for example, other slaveholders or other Nazis, since, given the prevailing norms in these societies, slaves and Jews are not considered relevant parties, or, indeed, "people." However, if we focus on the latter, then the thought is that the terms "other people" and "all concerned" is unqualified and should refer to all people who are affected by the action in question. This is so because the universal impartial spectator is a *notional* or *hypothetical* other person, any person, not a representative of a given society. Thus, the idea is that seeing, in an unqualified sense, that they are "but one of the multitude, in no respect better than any other in it" and judging "with the eyes of a third person […] who judges with impartiality" shows to both the

[12] Carson (1984, 51) argues that "if there are only some moral issues concerning which all ideal observers would have the same attitudes, then the ideal observer theory supports an intermediate view between extreme objectivism and extreme metaethical relativism. According to this intermediate view, there is an objectively correct view or judgment concerning *some*, but not all, moral questions." I am essentially offering a version of this type of view, albeit instead of referring to agreement or disagreement between different impartial spectators, I am making use of a universal impartial spectator who constraints society- and culture-relative virtues (or different localized impartial spectators).

slaveholder and the Nazi that they are mistaken in thinking that their own interests and points of view are more privileged than the slave's or the Jew's interests and points of view, respectively.[13]

The universal impartial spectator allows for the possibility of rightfully claiming that all the members of a given society, who abide by a practice approved from the standpoint of their societal impartial spectator, are mistaken. Nevertheless, there should also be a plausible explanation for how individuals within the society in question can transition from a societal impartial spectator to a universal one. The explanation I wish to offer, which I will exemplify with the case of slavery, builds on the ideas discussed above. Consider, first, knowledge of all the relevant non-normative facts. As Fleischacker (2011, 29) notes in his discussion of Smith's impartial spectator, better information about Africans and a better realization of how negative sentiments toward Africans served the interests of slave-owners might have helped adherents of the practice of slavery to realize that the practice is based on faulty information and has arisen to serve the interests of a certain group in their society. Second, the universal impartial spectator can develop in response to sympathy with marginalized individuals and groups of individuals. So, for example, if a slaveholder were to sympathize with his slaves—first by imagining himself in their situation but then by imagining them in their situation—he might come to realize that these people are part of

[13] One might ask about other attributes of the universal impartial spectator. For example, could he be cowardly, cruel, misanthropic, or dishonest? Do we also need to attribute benevolence to this spectator? In reply, recall that this is a notional person and, accordingly, Smith only attributes sufficient knowledge and impartiality, along with sympathy (TMS II.i.2.2 & VII.ii.1.49), to the impartial spectator. The thought that I have developed above is that these attributes would be sufficient to provide the requisite normative constraints on the virtues of a given society. It is worth noting, in regard to sympathy in particular, that Smith argues that just as people are pleased when others sympathize with them and are hurt by the lack of such sympathy, so people are pleased when they are able to sympathize with others and are hurt when they are unable to do so, even when we sympathize with painful sentiments (TMS I.i.2.1-6). See Fleischacker (2012, 301) for a defense of this point. Presumably the impartial spectator would incorporate these aspects of our human nature, and so the standpoint of this spectator would be one from which we are pleased when we are able to sympathize with others. I am grateful to an anonymous referee for pressing me clarify these points.

the "multitude." If we incorporate all of these suggestions into an explanation of the transition from a societal to a universal impartial spectator, we get the following plausible explanation: (a) A attains further information about and imagines himself in the situation of Bs, who are initially not considered A's equals from the point of view of A's societal impartial spectator; (b) these interactions start to alter A's societal impartial spectator in a way that takes Bs' interests into account; (c) A can now better imagine being a B in Bs' situation; (d) A sees Bs' interests and perspectives as equal to his own and views Bs as fellow human beings. The thought here is that as this process is repeated with different types of people, one's conscience can progressively be constrained by the *universal* impartial spectator. Of course, this is a psychological explanation that might not be true of all individuals, but it does demonstrate a path to the attainment of the standpoint of the universal impartial spectator. Accordingly, the account emphasizes the psychological features that make it possible to transcend the biases of one's culture and society.

Consider the case of John Newton (1725-1807), an English captain of slave ships who would go on to become an abolitionist. Although part of this transformation might be explained by his conversion to Christianity in 1748, his reflections on the slave trade and its moral status incorporate some of the key themes discussed above. In particular, being sufficiently informed about the circumstances and sympathizing with the slaves allowed Newton—and, per Newton, can allow others—to see the slaves as fellow human beings. Thus, in his essay "Thoughts upon the African Slave Trade" (1788), he notes how *misinformation* may lead people to justify the slave trade:

> Perhaps some hard-hearted pleader may suggest, that [the] treatment [of the female slaves] would indeed be cruel, in Europe; but the African Women are Negroes, Savages, who have no idea of the nicer sensations which obtain among civilized people. I dare contradict them in the strongest terms. I have lived long, and conversed much, amongst

> these supposed Savages. I have often slept in their towns, in a house filled with goods for trade, with no person in the house but myself, and with no other door than a mat; in that security, which no man in his senses would expect, in this civilized nation, especially in this metropolis, without the precaution of having strong doors, strongly locked and bolted. And with regard to the women, in Sherbro, where I was most acquainted, I have seen many instances of modesty, and even delicacy, which would not disgrace an English woman. [pp. 21-22]

Later in the text, Newton notes that he has "often, been gravely told, as a proof that the Africans, however hardly treated, deserve but little compassion, that they are a people so destitute of natural affection, that it is common, among them, for parents to sell their children, and children their parents." He responds that people who make such claims are "misinformed" and that he "never heard of one instance of either" (p. 31). Newton also demonstrates his *sympathy* with the slaves, understood in terms of taking their perspectives: "When a hundred and fifty or two hundred stout men, torn from their native land, many of whom never saw the sea, much less a ship, till a short space before they are embarked; who have, probably, the same natural prejudice against a white man, as we have against a black; and who often bring with them an apprehension that they are bought to be eaten: I say, when thus circumstanced, it is not to be expected that they will, tamely, resign themselves to their situation" (p. 14). He believes that sympathy can help others appreciate the wrongfulness of slavery. For example, after describing an incident in which a crewman threw the infant of a slave overboard because the infant was crying and disturbed his sleep, Newton adds that he is "persuaded, that every tender mother, who feasts her eyes and her mind, when she contemplates the infant in her arms, will commiserate the poor Africans" (p. 18). Another example:

> When the Women and Girls are taken on board a ship, naked, trembling, terrified, perhaps almost exhausted with cold, fatigue, and hunger, they are often exposed to the wanton rudeness of white Savages. The poor creatures cannot understand the language they hear, but the looks and manner of the speakers, are sufficiently intelligible. In imagination, the

> prey is divided, upon the spot, and only reserved till opportunity offers, Where resistance, or refusal, would be utterly in vain, even the sollicitation of consent is seldom thought of. But I forbear. — This is not a subject for declamation. Facts like these, so certain, and so numerous, speak for themselves. Surely, if the advocates for the Slave Trade attempt to plead for it, before the Wives and Daughters of our happy land, or before those who have Wives or Daughters of their own, they must lose their cause. [pp. 20-21]

While these examples do not demonstrate that he developed the universal impartial spectator, it is striking that Newton claims that he is "bound, in conscience, to take shame" in the fact that he was once part of a practice which "contradicts the feelings of humanity" (pp. 1-2). As I have argued, the impartial spectator can be understood as one's conscience, and the appeal to the "feelings of humanity" may suggest the type of sympathy that is felt from the standpoint of a "man in general."

3. The Impartial Spectator and Ideal Observer Theory

Smith's impartial-spectator account does not include the type of idealization that is inherent in Firth's ideal observer account,[14] but is it an ideal observer account of any kind? There are three questions that one could raise in this regard: (a) Does the impartial-spectator account provide conditions that are sufficiently specified and hypothetical? (b) Are the verdicts of the impartial spectator normative and universal? (c) Does the account avoid both vacuity and circularity? Let's commence with the first question. As the previous discussion suggested, the use of the standpoint of the impartial spectator is a *process* that is improved with additional experience and knowledge. Thus, Smith writes of the wise and virtuous man, for example, that "every day some feature [of the idea of exact propriety and perfection] is improved—every day some blemish is corrected"

[14] In this regard, see, for example, Campbell (1971, ch. 6), Griswold (1999, 144), & Sayre-McCord (2010).

(TMS VI.iii.25). In other words, impartial spectatorship takes the form of an open-ended process of revisions to our judgments based on new experiences and knowledge. So one might worry that Smith's emphasis on our moral judgments as refinable is in tension with an attempt to use the impartial spectator as a modest ideal observer theory, since it seems to get us further away from a set of defining features that characterize an ideal observer. However, as we saw, Smith does define the notion of impartiality inherent in the impartial-spectator account very clearly, namely, in terms of viewing the situation "from the place and with the eyes of a third person, who has no particular connexion" with the agents in question. And while it is true that the amount of knowledge required in order to make various judgments will vary as a function of the situation—Smith himself provides ample examples of this in TMS III.2.4—this is a strength rather than a weakness of the theory, since it demonstrates a nuanced sensitivity to the complexity of any given situation. Moreover, we are familiar with other accounts, for example, Bernard Williams's account of practical reasons, which utilize idealization but leave the specifications of the idealized conditions open-ended. In particular, Williams (1981) argued that *p* is a reason for A to φ only if *p* can motivate A to φ under suitable conditions, such as having no false beliefs and all the true beliefs *pertaining to the action* as well as deliberating correctly, which Williams left *open-ended*.[15]

One might further worry that the standpoint of the impartial spectator is too strongly grounded in *actual* human interactions and so is perhaps no more than the standpoint constituted by the

[15] One might object that the amount of information the impartial spectator needs in complicated cases—for example, a complicated political decision—might exceed ordinary human capacities. I concede that this might be the case. However, my proposed account would be, in this respect, in the same boat with many moral theories: it seems true of utilitarianism, for example, that the amount of information needed in order to ascertain the right course of action in complicated cases may exceed ordinary human capacities. I am grateful to anonymous referee for raising this worry.

'normal' reactions of a person within a certain society.[16] Accordingly, this standpoint cannot serve as the *hypothetical* suitable conditions from which moral judgments ought to be made. However, recall that while the impartial spectator grows out of actual interactions, Smith argues that we end up with an imagined "man in general" or "the representative of mankind," which, I argued, ought to be construed as a universal impartial spectator that can only be understood in hypothetical terms. Indeed, Smith puts a lot of emphasis on counterfactual reasoning in TMS, which allows us to free ourselves from the responses of actual people. For example, by imagining being in the victim's situation, we frequently experience resentment against injustice even when those suffering the injustice do not (TMS II.i.2.5).[17] Furthermore, the development of the universal impartial spectator is supposed to allow the standard of correctness set by the impartial spectator to potentially transcend the biases of the society which gave rise to it. In this regard, Smith himself not only notes that the standpoint of the impartial spectator can be used to *correct* the reactions of the actual people we encounter when those reactions are deemed inappropriate from this standpoint (TMS III.2.32 & VII.iii.3.9), but he also leaves open the possibility that no member of a given society has perfectly 'normal' reactions, for even the wise and virtuous man only approximates the impartial spectator's judgments (TMS III.3.25 & VI.iii.25). The key thought in this regard, which can be extracted from the discussion in section 2, is that the account provides hypothetical conditions—in the form of the standpoint of the universal impartial spectator—that guarantee the authoritativeness of an agent's (or agents') responses in constituting the standard in question, and,

[16] Campbell (1971, 145), for example, argues that the impartial spectator refers "to the normal reaction of a member of a particular group, or of a whole society, when he is in the position of observing the conduct of his fellows."
[17] See Schliesser (2017, 118-21) for an excellent discussion of counterfactual reasoning in Smith's account of sympathy.

if an actual agent or an actual community of agents are not under those conditions, their responses are not authoritative in setting this standard. However, the conditions can themselves be constructed from the psychology and interactions of actual human beings. In other words, facts about the morally appropriate and inappropriate are determined from hypothetical conditions that—while agents in a given society might have yet to attain them—can be constructed from those agents' shared experiences.

One might also worry that the verdicts of the impartial spectator are neither normative nor universal. In particular, one might worry that the impartial spectator is not a *normative* ideal, but is merely intended to be a sociological and psychological *explanation* of our moral capacities.[18] My argument that the impartial spectator provides hypothetical conditions which guarantee the authoritativeness of the responses of those adopting them already shows that the verdicts of the impartial spectator are normative. Here it is worth adding that Smith not only asserts that the sympathetic feelings of the impartial spectator are the "precise or distinct measure by which [the] fitness or propriety of affection can be ascertained" (TMS VII.ii.1.49), but also that the "wise and virtuous man" aspires "to assimilate his own character to this archetype of perfection," imitating "the work of a divine artist" (TMS VI.iii.25).[19] One might further worry that Smith's impartial spectator account is too context-dependent to provide universal moral principles. The preceding discussion regarding the universal impartial spectator suggests that universal principles are at least possible in the impartial spectator framework, since *all* the members of a society may be wrong

[18] See, in particular, Campbell (1971, 145) & Raphael (2007, 47-48).

[19] While Smith does write in a footnote that "the present inquiry is not concerning a matter of right […] but concerning a matter of fact" (TMS II.i.5.10), he is clearly looking for a standard that determines what is "fit, and right, and proper to be done" (TMS III.5.5).

regarding a certain moral judgment. Accordingly, Smith writes of the infanticide of ancient Greece, for example, that it "was permitted from views of remote interest or conveniency, which could by no means excuse it. Uninterrupted custom had by this time so thoroughly authorized the practice, that […] the loose maxims of the world tolerated this barbarous prerogative" (TMS V.2.15). One way in which we can come to formulate universal moral principles is via induction. As Smith argues, "the general rules of morality" are "ultimately founded upon experience of what, in particular instances, our moral faculties, our natural sense of merit and propriety, approve or disapprove of," that is, "by finding from experience that all actions of a certain kind […] are approved or disapproved of" (TMS III.4.8). This generalization is made from our sentiments as experienced from the standpoint of the impartial spectator: this is suggested in Smith's talk of "our natural sense," which he often uses to refer to sentiments felt from the standpoint of the impartial spectator. [20] But recall that while Smith's impartial spectator account allows for the possibility of transcending the biases of a society, and for the formulation of universal principles, it can also give normative standing—in a way that Firth's ideal observer theory cannot easily accommodate—to some of the particular virtues of a given society. This is a substantial advantage over Firth's theory, since it allows for a more fine-grained normative response to the specifics of a certain society.[21]

[20] One plausible possibility is that those adopting the standpoint of the impartial spectator share their verdicts, and/or their verdicts can be compared to each other by others, in order to formulate universal principles. When such rules have been established, "we frequently appeal to them as to the standards of judgment, in debating concerning the degree of praise or blame that is due to certain actions"; they correct "the misrepresentations of self-love concerning what is fit and proper to be done" and are "commonly cited as the ultimate foundations of what is just and unjust in human conduct" (TMS III.4.11-12). For further discussion of these points, and related issues, see Ben-Moshe (2021).

[21] Even in the case of infanticide, Smith notes that there is a *limited* context-dependence: "[T]his practice prevails among all savage nations; and in that rudest and lowest state of society it is undoubtedly more pardonable than in any other. The extreme indigence of a savage is often such that […] it is frequently impossible for him to support both himself and his child. We cannot wonder, therefore, that in this case he should abandon it" (TMS V.2.15). For an excellent discussion of the tension between universalism and relativism in Smith's moral philosophy, see Fleischacker (2011).

Finally, one might worry that the account is vacuous or circular. First, given the framework's anti-realist commitments, a metaphysically constitutive (rather than epistemological) impartial spectator cannot make decisions on the basis of mind-independent moral facts. This can lead to the following worry: either one makes moral judgments according to local standards, but fails to transcend the biases of those standards; or one steps back from all such standards, but lacks a basis for making moral judgments at all and thus the standard that is attained is vacuous. The universal impartial spectator offers a middle road: instead of completely stepping away from all local standards, it acts as a normative constraint on such standards; more specifically, it allows many moral judgments to be made according to local standards, while holding those standards accountable to the ideal that everyone's interests and points of view should be taken into account. Understood this way, the standard of the universal impartial spectator can both transcend contingent local standards and not be vacuous. However, and second, one might worry that the account is viciously circular, because equating the standard of moral judgment with the universal impartial spectator merely shows that we have a *pre-commitment* to impartiality when assessing things from a moral point of view. I wish to bite the bullet and argue that the impartial spectator does build on our pre-commitment to associating morality with impartiality. As Taliaferro (1988, 127) observed, when presenting his ideal observer theory, "in developing a proper characterization of the ideal moral point of view we can, I think, do no better than reflect upon our ordinary moral judgments, reflect on the pre-philosophical data, and guide our philosophical reflection by our best intuitions on the nature of morality." Now, we cannot assume from the outset, on pain of circularity, that the impartiality in question necessarily encompasses all human beings. However, all I am assuming is that, as a matter of fact, we associate morality with impartiality, while the

scope of this impartiality is left open-ended. The thought is that even if they are not committed to taking the interests and points of view of all human beings into account, even the slaveholder and the Nazi are generally committed to the idea that when weighing the competing interests of fellow slaveholders or Nazis, their responses should be impartial in order to be correct. The impartial spectator account presupposes only this modest commitment to impartiality in moral reasoning; the fact that the impartial spectator comes to encompass all people is a gradual achievement, not a presupposition.[22]

4. The Motivational Challenge

In the remainder of the paper, I wish to address three challenges to an ideal observer theory. The first challenge is a motivational challenge. Nick Zangwill (2003, 285-6) has argued that non-normative response-dependence theories cannot tell a plausible story about why we pursue what we believe to be morally good. In particular, he argues that, according to such theories, when we are motivated to pursue morally good actions, we are aiming to acquire the response-dependent property of being disposed to elicit approval in certain people (either others or ourselves) under certain conditions. However, surely it is not the case, Zangwill argues, that our reason for performing morally good actions is always that we want either others or ourselves to approve of our actions, or to be disposed to approve of them. Rather, it seems that we sometimes perform certain actions *because* they are the right thing to do. Smith's impartial-spectator account can offer

[22] These observations about circularity also pertain to the inclusion of sympathy as a characteristic of the impartial spectator: according to sentimentalist reasoning, this inclusion builds on our pre-commitment to associating certain features of human nature (sympathy) with morality in general and with the formation of moral judgments in particular.

a satisfactory answer to this challenge via the desire to be worthy of approval.[23] In particular, Smith makes a clear distinction between a desire for mere approval and a desire to be worthy of approval. The difference between the two is that "praise and blame express what actually are; praiseworthiness and blameworthiness, what naturally ought to be the sentiments of other people with regard to our character and conduct." Therefore, as Smith continues, while the love of praise is "the desire of obtaining the favourable sentiments of our brethren," the love of praiseworthiness is "the desire of rendering ourselves the proper objects of those sentiments" (TMS III.2.25). Thus, the desire to be worthy of approval is a desire to be praised by the right type of agent in the right type of way. And it is ultimately the (universal) impartial spectator who can reliably praise us for doing the right thing, since the standpoint of this spectator constitutes the morally appropriate and inappropriate. Indeed, Smith clarifies that while agents might initially wonder whether they are worthy of the approval of their fellow human beings, only the impartial spectator can satisfy this desire: while the jurisdiction of people in society is founded "in the desire of actual praise, and in the aversion to actual blame," the jurisdiction of the impartial spectator is founded "in the desire of praise-worthiness, and in the aversion to blame-worthiness" (TMS III.2.32). Therefore, when we act for moral reasons, we are not doing so because we want mere approval from others or from ourselves. Rather, we are motivated by a desire for *warranted* approval, and, in doing so, seek this

[23] Zangwill (2003, 289) mentions the desire to be worthy of approval as a candidate that meets the motivational challenge, but associates it with *normative* response-dependence theories of value and argues that they are vacuous.

type of approval from the sort of agent who can provide it, namely, the (universal) impartial spectator.[24]

The existence of a desire to be worthy of approval and its centrality in human psychology is highly plausible. One can provide the following account of how such a desire might develop in two key stages. First, a desire for *weak approvability* is formed: we desire to be praised only for actions that we performed and for the reasons for which we actually performed those actions. This desire cannot be satisfied by praise for φ-ing if one did not φ or by praise for φ-ing for reason *p* if one did φ, but did it for some reason other than *p*. To be praiseworthy in this sense is to have done the thing for which praise is being offered. This is a very commonplace phenomenon. Consider a child who feels satisfied when his mother praises a drawing which she thinks he has produced. As it turns out, it is the child's brother who actually drew the picture. While the child might initially be satisfied by his mother's approval, he will probably come to be dissatisfied by this form of approval precisely because his mother is wrong about the relevant facts (the identity of the painter). It is the development of the desire for weak approvability that accounts for the interest that agents have in correcting biases that are the result of others not being fully informed about the facts. Second, a desire for *strong approvability* is formed: we desire to be praised only by agents who are well-suited to praise us. This desire cannot be satisfied by praise for φ-ing if one knows that agent A is not well-suited to praise one for φ-ing. Given Smith's account, to be praiseworthy in

[24] Kawall (2004) argues that Zangwill overlooks the fact that (a) there are different kinds of approval, and (b) the nature of the individuals whose moral approval we seek is important. In connection with (b), I am indeed arguing that approval from the impartial spectator is different from approval from actual spectators; the former, but not the latter, has the requisite normative authority. In connection with (a), instead of focusing on different kinds of approval in Kawall's sense—he asks us to compare approval towards a moral hero, a beautiful painting, or the skillful tactics of an opponent in a game—I am arguing that the desire to be *worthy* of approval is different from the desire for approval.

this sense is to be praised by people who do not have a personal stake in praising us. It is also a commonplace experience that can be exemplified by returning to our example: once our child is satisfied with his mother's approval of his drawings only when she in fact knows that these are his drawings, he might, at some point, not be fully satisfied by her praise precisely because she has a personal stake in the circumstances; thus, the child might seek a more impartial judge to comment on the quality of his drawings, a judge who does not have a personal stake in the circumstances and who is thus better-suited to assess the merits of the drawing qua drawing. As this process continues, people will tend to seek approval from spectators who are fully informed about the situation *and* do not have a personal stake in the circumstances. It is thus plausible that people will ultimately seek approval from an impartial spectator, who exemplifies these characteristics. At this point, the desire to be worthy of approval can be satisfied if we know that the impartial spectator *would* approve of our actions, even if such approval is not provided by people around us (TMS III.2.5).[25]

The culmination of the process described above might get us only to a *societal* impartial spectator: after all, the examples I discussed do not yet demonstrate transcendence of a given society's biases. In section 2, I described how a universal impartial spectator can arise when people

[25] These developmental stages are not articulated by Smith, but are rather my attempt to demonstrate that such a desire can develop over time in individuals. Smith, for his part, thought that the desire to be worthy of approval is "natural," that is, hard-wired into our psychology: "Man naturally desires […] to be lovely; or to be that thing which is the natural and proper object of love. He naturally dreads […] to be hateful; or to be that thing which is the natural and proper object of hatred" (TMS III.2.1). Moreover, Smith notes that the desire to be worthy of approval is not derived from the desire for approval, and that the two desires are "in many respects, distinct and independent of one another" (TMS III.2.2). However, the desire to be worthy of approval cannot be *fully* formed at the outset, since the standard of the impartial spectator—the standpoint that determines what is worthy of approval—is not yet in place. The distinction between the two developmental stages, along with the accompanying examples, is borrowed from Ben-Moshe (2020a, 1079 & 1086).

attain further information and sympathize with certain types of individuals. This process would probably also involve judgments about oneself in the following way: one would start to realize that potentially most, or even all, of the members of one's society might not be sufficiently informed about certain situations and/or have a personal stake in the circumstances in question. As this happens, one's desire to be worthy of approval would be further refined so that this desire is no longer fully satisfied by approval from the societal impartial spectator, who is representative of these people. Rather, one would also seek approval from the universal impartial spectator. (Or, to put the point in a more psychologically plausible way, one would seek approval from one's societal impartial spectator, but only insofar as the latter is constrained by the universal impartial spectator.) Put this way, the account might seem to force us into the following dilemma. According to the first horn, if we rely on the standard of the universal impartial spectator as defining the contents of the desire to be worthy of approval, we risk circularity. According to the second horn, we do not rely on the standard of the impartial spectator as defining the contents of the desire; however, since the desire to be worthy of approval is morally relevant only if it is conducive to what is morally appropriate—which is determined from the standpoint of the impartial spectator—we would now open up the possibility of desiring to be to be worthy of approval for doing what turned out to be wrong. The first horn of the dilemma is avoided via the developmental story of the desire to be worthy of approval, since the account does not assume the standard of the universal impartial spectator from the start; rather, both standard and desire develop in interaction with one another. So there is no standing desire to be worthy of approval whose objects are defined, from the get-go, by the standard of the universal impartial spectator. Regarding the second horn, a two-part reply can be offered. First, the desire to be worthy of approval, when *fully* formed, is in fact

conducive to what is morally appropriate, because the universal impartial spectator, which fully satisfies it, determines what is ultimately worthy of approval. Second, while it is true that until this desire reaches its final stage of development, agents may desire to be worthy of approval for doing what turned out to be wrong, the motivational challenge is still met, since Zangwill (2003, 286 & 288), whose focus is on "our folk conception of moral motivation," saw the challenge as pertaining to the pursuit of "what we *believe* to be morally right" (emphasis added).[26] In my proposed account, agents do act on what they believe to be worthy of approval, even in the desire's initial form.[27]

5. The Ad-Hocness Challenge

The second challenge to an ideal observer theory that I wish to discuss is an ad-hocness challenge. David Enoch (2005, 761-765) has argued that idealizing theorists are not likely to be able to motivate the idealization that they employ and so the idealization is likely to remain objectionably ad hoc. In particular, Enoch argues that response-dependence views of the normative cannot consistently employ the natural rationale for idealization—according to which responses under

[26] Zangwill's mention of our folk conception of moral motivation suggests that he believes that the motivational challenge can be met if the motivation in question refers to a *normative* concept such as duty or rightness or worthiness, rather than, for example, mere approval or self-interest. He does not understand the motivational challenge as being focused on the question of whether the motivation is necessarily successful in attaining what is morally appropriate.

[27] It is worth noting that I have not conclusively shown that most human beings desire to be worthy of approval, even in the modest sense of what I have called "weak approvability." Providing empirical studies confirming this hypothesis, insofar as they exist, is, alas, beyond the scope of this paper. Moreover, even if the hypothesis is generally true, it is still possible that there are individuals who do not desire to be worthy of approval at all. However, as I argue elsewhere, apart from the fact that it is questionable whether there are any non-disabled human beings above a certain age who do not desire to be worthy of approval at all, if there are such agents, we may not want to say that they are distinctively human and so we would not know what to make of the application of moral standards to them. Indeed, Smith argues that human beings have developed the desire to be worthy of approval, and not only the desire for mere approval, as the latter "would not alone have rendered [them] fit for the society for which [they were] made"; rather, it is the desire to be worthy of approval that is "necessary" to make human beings "really fit" for society (TMS III.2.7). Therefore, the desire to be worthy of approval is needed in order to make us the type of social beings that we are, and our social systems, or perhaps even our species, are structured so as to instill such a desire (Ben-Moshe 2020a, 1084).

idealized conditions serve as *evidence of* or *reliably track* mind-independent facts (for example, a watch reading under suitable conditions reliably tracks the time, which is independent of it)—since this rationale would undermine their claim that it is the responses of agents under certain conditions that *constitute* the facts in question. Of course, it might be the case that this "natural rationale," which applies to physical properties, does not apply to normative properties. Accordingly, Enoch (2005, 769-70) concedes that one of the rationales that idealizing theorists of normative facts could use to motivate their theory in a way that is not objectionably ad hoc is to argue that idealization is warranted given our justificatory practices: we believe that we are fallible in our normative judgments and that there is room for genuine normative advice—for example, if one maintains that lifelong solitude is of value but is shown not to have a good appreciation of what such a life consists of, then one has been discredited as a competent judge of the value claim—and so we also believe that from our practices of justifying normative claims a rationale for idealization can be extracted. These theorists would be appealing to an inference to the best explanation in the sense that the rationale for the idealization is that it best explains important aspects of the relevant justificatory practice(s). However, Enoch (2005, 776) argues that this rationale, namely, appealing to our justificatory practices, will not do, because our moral discourse incorporates an objective purport, which can be captured by realism but not by response-dependence views of the normative. He concludes that it is not the case that the latter views best explain our justificatory practices.

The Smithian account suggests the following when it comes to our justificatory practices regarding a spectator-based account of the morally appropriate and inappropriate: if we endorse a spectator-based account of morality, and if we come to realize that actual spectators might be fallible in their normative judgments about our actions, then we should also realize that from our

practices of justifying claims about moral propriety and impropriety a rationale for idealization can be extracted. In particular: (a) actors come to realize that the spectators who judge them are fallible, because they are not informed about the non-normative facts and/or have a personal stake in the circumstances; (b) a way of fixing such deficiencies is by seeking approval from spectators who are sufficiently informed and who do not have a personal stake in the circumstances; (c) these justificatory practices suggest that a promising way of attaining approval from the right type of spectator is by imagining a modestly idealized spectator, namely, the impartial spectator (first a societal impartial spectator and then a universal impartial spectator); (d) since people's patterns of approval and disapproval *constitute* our initial assessment of propriety and merit—recall Smith's model of approbation—it is warranted to assume that patterns of approval and disapproval of agents under the relevant idealized conditions also *constitute*, and are not evidence of, what is in fact morally appropriate and inappropriate. Thus, what best explains our justificatory practices is the fact that the standard of moral judgment is constituted by our responses under idealized conditions. Points (a) through (c) show why we have reason to think that we ought to adopt the standpoint of an impartial spectator: actual spectators assess us and make judgments about the propriety and impropriety of our conduct; we come to realize that these initial assessments may be distorted; and so we imagine a spectator who would be comparatively free of these distortions. Given the constitutive role that our patterns of approval have in our initial assessment of propriety and merit, point (d) provides us with a reason to think that the standpoint of the (universal) impartial spectator constitutes the morally appropriate and inappropriate. Taken together, (a) through (d) show that we have reason to think that we ought to adopt the standpoint of the

(universal) impartial spectator and that this standpoint constitutes the morally appropriate and inappropriate.[28]

A realist might still argue that while there are reasons to think that we ought to adopt the standpoint of the (universal) impartial spectator and that this standpoint constitutes the morally appropriate and inappropriate, there are even more compelling reasons to think that our justificatory practices are committed to realism, because of the objectivity that is lacking in idealizing theories. While I cannot conclusively settle the issue in this paper, I believe that our justificatory practices commit us to two key requirements, both of which a modest ideal observer theory of the kind that I have presented can account for. First, we want the standard of correctness of moral judgment to be more robust than a subjectivist or relativist standard, that is, we want to be able to say that this standard is not dependent on what each one of us, or even a collection of us, approves and disapproves of. Second, and relatedly, we want to make intelligible the idea that we might all be wrong, and perhaps have always been wrong, when making moral judgments.[29]

[28] In an earlier paper, I made similar observations about a reconstruction of Hume's general point of view (Ben-Moshe 2020b, 445-446). It is worth noting that I focus primarily on the idea that our justificatory practices vindicate idealizing, since I believe that this is the most promising strategy for motivating ideal observer theory. However, there may be other alternatives. For example, in a paper that criticizes Enoch's position, Sobel (2009, 343) argues that the most obvious rationale for idealization is to provide the agent "with a more accurate understanding of what the option she is considering would really be like" (though Sobel's focus in the paper is on subjectivist (desire-based) accounts of well-being).

[29] Realists like Enoch would argue that our justificatory practices are committed to more than these requirements, namely, to (a) truth and hence also to (b) mind-independent normative facts. My aim is not to take a position about (a) or about whether (a) entails (b). Rather, my aim in the discussion above is to show that the constitutive role of the idealization in Smith's account is well-motivated, and, furthermore, that two key requirements of our justificatory practices are satisfied. Nevertheless, it is worth noting that while Smith is a meta-ethical sentimentalist—he notes, for example, that "the first perceptions of right and wrong […] cannot be the object of reason, but of immediate sense and feeling" (TMS VII.iii.2.7)—his moral philosophy is not hostile to the idea that moral judgments can express genuine beliefs and hence be true or false. Thus, Smith's moral philosophy is compatible with the aspirations of Firth's ideal observer theory, according to which an ideal observer's reactions determine, as noted earlier, whether an ethical judgment is true or false.

As argued, the impartial spectator framework specifies hypothetical conditions from which the objects of our patterns of approval merit that approval. When the universal impartial spectator has been attained, the standard of correctness in question does not exclusively rely on the patterns of approval of actual individuals or, indeed, on those of collections of individuals. Accordingly, the proposed framework provides a standard of correctness that can transcend the patterns of approval, and hence the biases, of individuals and entire communities. Therefore, we might all be, and/or have been, wrong when making a certain moral judgment, if it were the case that the judgment would not be endorsed by the universal impartial spectator. Of course, given the fact that the standpoint of the impartial spectator grows out of antecedent interactions with others, it is not likely to be the case that *all* of one's moral judgments are wrong—indeed, the universal impartial spectator may endorse many of the local judgments that are made within a given society or culture—but this observation is probably true of any account of moral judgment that is sensitive to the texture of human life and human sensibilities. Nevertheless, the impartial spectator account does offer a sound basis for the claim that any person or group of people in any era may be wrong about a given moral judgment, for example, those pertaining to the moral permissibility of slavery or genocide.

6. The Standards-for-Standards Challenge

The third challenge to an ideal observer theory that I wish to discuss is a standards-for-standards challenge. Geoffrey Sayre-McCord (2010, 137-8) has argued, in connection with Smith's (and Hume's) project, that "there is a real possibility that once we uncover and examine our standards,

we'll discover that, by our own lights, they don't stand up to scrutiny. In those cases, we will then have found reason to change them. Alternatively, though, we might discover that our standards, once examined and understood, actually withstand the test well." In other words, Sayre-McCord argues that if a standard we have been relying on in making our moral judgments fails, upon further reflection, to meet our standards, then we have reason to think the standard is wrong as a standard for what is morally appropriate and inappropriate. He then applies this general idea to Smith's account of the impartial spectator, arguing that Smith uses this reflective endorsement test and that, when doing so, he relies on the very standard of moral judgment that he has defended, namely, the standard of the impartial spectator. A similar point is made by Samuel Fleischacker (2013), who insists that "Smith essentially provides what Christine Korsgaard calls a 'reflective endorsement' argument" and that "a good way to read TMS is to see Smith as demonstrating, to an impartial spectator in a moment of reflection, that the impartial spectator we use in the course of action operates in a reasonable and noble way." Now, although Smith actually makes little use of the reflective endorsement test in TMS,[30] a defense, upon further reflection, of the standard of moral judgment—that is, of the standard of the impartial spectator—seems desirable for the following reason: while Smith repeatedly refers to the impartial spectator as "ideal" and even as a

[30] Smith briefly discusses reflective endorsement in connection with Hutcheson's position, with which Smith disagrees, according to which moral predicates do not apply to the moral sense (for the same reason that sensory predicates do not apply to the senses—sight, for example, cannot be said to be black or white). In this regard, he merely argues that we can use the standpoint of the impartial spectator in order to reflect and pass judgment on the moral sentiments and faculties of *other spectators*. In fact, we have the ability to override the sympathy we feel towards the actor and assess the moral sentiments and faculties of the spectator judging the actor (TMS VII.iii.3.8-9). And more generally: "Correct moral sentiments […] naturally appear in some degree laudable and morally good" (TMS VII.iii.3.10). Therefore if, upon reflection from the standpoint of the impartial spectator, certain sentiments do not appear laudable and morally good, we have good reason to think that they are incorrect and cannot form the basis for moral judgments. In other words, Smith does not use reflection in order to test and ratify the standard of the impartial spectator itself; rather, he argues that, using the standpoint of the impartial spectator, we can reflect on the moral faculties of spectators and, indeed, on sentiments more generally.

"demigod,"[31] he also emphasizes the *human* nature of the impartial spectator, who is "of mortal extraction" (TMS III.2.32). He clarifies that he is not examining "upon what principles a perfect being would approve of the punishment of bad actions; but upon what principles so weak and imperfect a creature as man actually and in fact approves of it" (TMS II.i.5.10). And when discussing the influence of fortune on our judgments of others, Smith notes that this "irregularity of sentiment" is not felt only by "those who are immediately affected by the consequences of any action," but "in some measure, even by the impartial spectator" (TMS II.iii.2.1-2; see also TMS VI.iii.30). Thus, if the standard of the impartial spectator is, in fact, a 'defective' one, would we not need to defend the appropriateness of this standard upon reflection?

I wish to bring out a problem inherent in the attempt to use the reflective endorsement test in Smith's account, and then to generalize the problem for idealizing theorists. The problem—or standards-for-standards challenge—lies in trying to identify the standard that governs reflection on the standard of moral judgment; this forces us into the following dilemma, which is discussed by Sayre-McCord (2013, 233) in a later paper: According to the first horn of the dilemma, we should rely on the standard of the impartial spectator in reflecting upon this very standard. Put more precisely, we need to check whether the impartial spectator would approve of his own patterns of approval. However, recall that, according to Smith's model of approbation, when we recognize that there is concordance between our sympathetic passion and the original passion of the agent in question, a sentiment of approval arises. Therefore, the impartial spectator will necessarily approve of his own patterns of approval, for he will recognize, upon reflection, that he

[31] Regarding "ideal," see TMS III.3.26-29, III.3.38, & III.4.4; regarding "demigod," see TMS III.2.32, VI.iii.18, & VI.iii.25.

would have precisely the reactions which he does in fact have. Hence, the reflective test on this first alternative is trivially satisfied. According to the second horn of the dilemma, we require the approval of a second spectator, who is different from the impartial spectator, in order to defend the impartial spectator's patterns of approval. While this would make the reflective test nontrivial, it raises a new worry: Why do one spectator's patterns of approval set the standard of moral judgment, while the second spectator's patterns of approval set the standard employed in reflecting on the standard of moral judgment? Why can't it be the case that one spectator plays both roles? One way to avoid this dilemma, according to Sayre-McCord's (2013, 234) later paper, is to change the question: instead of asking whether the impartial spectator approves of his patterns of approval, we should ask whether the impartial spectator approves of the fact that *we* rely on the deliverances of the impartial spectator in making moral judgments. Sayre-McCord argues that this would constitute a nontrivial test, for Smith's account of approbation does not ensure that the impartial spectator will approve of us using the spectator's reactions as the standard for moral judgment. However, this merely pushes the worry one step back: when the impartial spectator approves or disapproves of the fact that we rely on the deliverances of the impartial spectator in making moral judgments, he will need to rely on some standard in reaching his verdict. Once again, it will either be the case that (a) the reflective endorsement test is trivially satisfied, because there is no conceivable reason for the impartial spectator not to approve, upon reflection, of us relying on his own deliverances in making moral judgments;[32] or (b) the impartial spectator will need to employ

[32] It seems that Sayre-McCord does not think that there is triviality involved here, because, contrary to the impartial spectator approving his own patterns of approval, it is logically possible that he would not approve of us using this standard. However, surely triviality is not merely a function of lack of alternative logical possibilities: if an agent needs a reason to rule out *p*—as the impartial spectator would need a reason, on pain of arbitrariness, to make the

a different standard to determine whether he would approve of us using the standard of the impartial spectator, which will, again, raise questions about the relation between the two standards.

If reflective endorsement is a normative test for the appropriateness of (modest and non-modest) idealized response-dependent standards, then the standards-for-standards challenge might generalize to potentially all ideal observer theories. First, we need a standard to govern reflection, since reflection can be biased, and so one might wonder whether subjecting our moral standards to critical scrutiny and finding that they pass reflection means that we have reason to endorse them.[33] Second, the standard that governs reflection would need to be different from the standard of moral judgment, on pain of triviality, but non-arbitrarily related to the first standard, so that we have a compelling explanation for the relations between the two standards. In order for the requisite standard to be not only non-trivial but also non-arbitrary, I wish to suggest that in the case of a modest ideal observer, such as Smith's impartial spectator, the standard used in reflection should pertain to the reason we have for preferring this modest observer to Firth's observer. In particular, we should use a standard of *accessibility* when reflecting on the standard of moral judgment in the following sense: is it the case that, contrary to Firth's account, we could identify the best version of *our* ideal ethical selves in the modestly idealized observer? Smith, for his part, argues that we often do identify ourselves with and become the impartial spectator, who, as noted, he refers to as "conscience."[34] This second-order standard gives us a good reason to believe that it is appropriate

decision regarding his disapproving of us using this standard—but there is no conceivable reason for him to do so, it would seem that *p* is as trivial as it would be if there were no alternative logical possibilities.

[33] In this regard, see in particular Kornblith (2012).

[34] Smith does argue that there are some situations in which, while we can imagine what the impartial spectator would approve of and closely approximate the impartial spectator's judgment, we will do so only imperfectly (TMS I.i.5.8). Moreover, as noted earlier, Smith suggests that even the wise and virtuous man can only approximate the impartial

to use the standard of moral judgment in question: if we cannot identify the best version of *our* ideal ethical selves in the idealized spectator, then it is not clear why this standard should bind or motivate us. Accordingly, the second-order standard is non-arbitrarily related to the first-order standard of moral judgment by providing a link between the latter standard and our ability to recognize its normative authority over us and to be motivated to act in accordance with it. Of course, more needs to be said about the nature of this link and about the scope of "our" in the phrase "our ideal selves." I will do so by explaining why the patterns of approval of the (universal) impartial spectator, qua ideal self, set the standard of moral judgment, while the actual self's patterns of approval set the standard used in reflecting on the standard of moral judgment. That is, I will explain why *we*—and not the impartial spectator, per Sayre-McCord's revised view—approve of the fact that we rely on the deliverances of the impartial spectator in making moral judgments.

It is helpful to think of the modest ideal observer in terms of an ideal *self*: there is a certain distance between the judgments of the agent's actual self and those of his ideal self (the modest ideal observer)—the latter is under suitable conditions and determines the correct normative claims for the former. However, given the relatively modest idealization in question, the agent's actual self can not only recognize that the ideal self is better situated to determine what ought to be done, and hence that the ideal self has the requisite authority over it, but the actual self could also recognize itself in the ideal self. The nature of this "could" should be cashed out as follows. First, even the idealization inherent in the universal impartial spectator is attainable for most people, if

spectator's judgments. Nevertheless, given Smith's other views about the impartial spectator, it is clear that most of us can become the impartial spectator to the degree needed for this ideal observer to count as our ideal self.

they are open to obtaining knowledge of the relevant non-normative facts, sympathizing with others, and using their imagination. Second, most human beings desire, at least to some extent, to be worthy of approval, and so they could deliberate from this desire—and its vicissitudes, as discussed in section 4—to the conclusion that the standpoint of the universal impartial spectator, which determines what is worthy of approval, has normative authority over them.[35] Combining both these points, the key thought is that most slaveholders and Nazis, for example, could use their imaginative capacities to sufficiently approximate the standpoint of the universal impartial spectator and deliberate from their desire to be worthy of approval to the conclusion that this standpoint has normative authority over them, *even if* they might choose not to do so. In other words, I am not claiming that most human beings will in fact adopt the standpoint of the universal impartial spectator, but rather that, given their initial motivations, as well as their imaginative and sympathetic capabilities, they could do so; accordingly, this standard does in fact have normative authority over them and could motivate them to action. Finally, note that since even the universal impartial spectator respects important societal and cultural differences, we do not all have *identical* ideal ethical selves: the ideal ethical self of a person living in society A at t1 might not be identical to that of a person living in society B at t2, since certain virtues in their respective societies would be factored into the verdicts of the universal impartial spectator. Rather, the *normative constraints* that are placed on people's ideal ethical selves, and which are determined by the universal impartial spectator, are identical. So the slaveholder's and the Nazi's ideal selves would retain many of the

[35] I am following in part the later Williams (1995, 35; 2001, 91) when he argued that *p* is a reason for A to ϕ only if A *could* reach the conclusion that he should ϕ by a sound deliberative route from his motivations. For a discussion of this point in connection with Smith's impartial spectator account—albeit one that does not differentiate between a societal and a universal impartial spectator—see Ben-Moshe (2020a, 1079-80).

virtues considered worthy of approval in nineteenth-century America or in 1930s Germany, respectively. However, these selves would be constrained by norms that do not permit slavery and genocide, since the interests and perspectives of all concerned are to be taken into account.

Now, one might worry that while I have provided a meta-standard to govern reflection, we would also need a meta-meta-standard to evaluate the proposed meta-standard, and then a meta-meta-meta standard and so on ad infinitum. Three replies can be provided to this worry. First, reflection on the standard of moral judgment does not determine the correctness of the moral judgments themselves; the idealized conditions in question, even if relatively modest, are supposed to guarantee that correctness. The reflective endorsement test is supposed to reassure us that the standard of moral judgment is an appropriate standard for our moral judgments. As Sayre-McCord (2013, 235) notes, "if it is our reliance on the standard that is up for evaluation, we will not be concerned with showing that what garners approval merits that approval," but rather "with showing that it is morally good (or appropriate, or justified) to use the fact that something garners approval (or disapproval) from a privileged point of view as the standard for our judgments." Put this way, the standard I have proposed, namely, the standard of accessibility, is supposed to provide us with *sufficient* reason to believe that the standard of the impartial spectator, even though 'human' in nature, is an appropriate standard of moral judgment. Moreover, this standard is non-arbitrarily related to the standard of moral judgment. Second, we could build on Frankfurt's (1971, 16) line of defense, when he considered the appeal to volitions of a higher order than the second, but dismissed the possibility by arguing that "when a person identifies himself decisively with one of his first-order desires, this commitment 'resounds' throughout the potentially endless array of higher orders." One could make an analogous case for the standard that governs reflection on the

standard of moral judgment: insofar as one not only identifies the standard of the impartial spectator with one's ideal self but also identifies with one's ideal ethical self, this identification "resounds" throughout the potentially endless array of higher-order standards (or, if intelligible, ideal ethical selves). The parenthetical remark leads me to the third, and perhaps most important, point: it is far from obvious that any higher standard of reflection, that is, any higher standard beyond an appeal to our *ideal* ethical self, would be intelligible to us, since it would transcend the imaginative lives of beings such as ourselves; or, alternatively, even if such a standard were intelligible, it would not have any resonance for us, since it would be detached from our human sensibilities. This appeal to the limits of our psychology, rather than to what is logically possible, is especially warranted in a theory in which the normative is tightly hooked to the psychological.[36]

7. Conclusion

I have built on Adam Smith's account of the impartial spectator in order to offer an adequate modest ideal observer theory. I have also argued that the idealized observer in the proposed account is related to our motivations in the right way, that the account is well-motivated, and that the standard of this idealized observer can be defended upon further reflection using a non-arbitrary second-order standard. Furthermore, I have argued that while Firth's ideal observer might be regarded as a *prima facie* 'better' idealized response-dependent standard, it is a standard of moral judgment that is detached from our human sensibilities and so raises obvious epistemic and practical concerns; to use the terminology from my discussion of the standards-for-standards

[36] I am grateful to an anonymous referee for urging me to clarify these points.

challenge, human beings could not identify with Firth's ideal observer as their ideal ethical self. Indeed, given my discussion of the motivational and ad-hocness challenges, we can add the following observations regarding Firth's standard. First, given the standard's detachment from our human sensibilities, it is far from clear that acting in accordance with the verdicts of this type of observer would satisfy our desire to be worthy of approval, for there is no reason to assume that this desire, which is a *social* desire at base, would be satisfied by approval from a spectator who is not recognizably human. Second, it is far from clear that our justificatory practices warrant the type of idealization inherent in Firth's account, for, contrary to the impartial spectator, Firth's ideal observer does not grow out of our interactions with actual people. So while the Smithian account does not include the idealization of Firth's theory, it is an ideal observer account that the naturalist anti-realist should embrace, for, to paraphrase a point I made in connection with the second-order standard used in reflection, there is no higher standard for the correctness of moral judgment which would fare better: any higher standard, above and beyond the one set by the impartial spectator, would be unintelligible to us, since it would transcend the imaginative lives of beings such as ourselves; or, alternatively, even if such a standard were intelligible, it would not have any resonance for us, since it would be detached from our human sensibilities. Of course, realists and rationalists will argue that their accounts can provide a better standard for moral judgment. My aim was to appeal primarily to those naturalist anti-realists who find ideal observer theory appealing, not to settle the meta-ethical debate regarding which meta-ethical position is, all things considered, most plausible. However, as noted, my account has two key upshots: (a) it provides a standard of moral judgment that is more robust than a subjectivist or relativist standard; (b) it

makes intelligible the idea that we might all be, and/or have been, wrong when making moral judgments.

Acknowledgments I would like to thank audiences at the APA Central Division Meeting (2020) and the Central States Philosophy Association Meeting (2019) for their feedback. I am especially grateful to Derrick Baker, Samuel Fleischacker, Ben Miller, James Rowe, Anthony Rudd, Karsten Stueber, David Sussman, and Alyssa Walker for detailed and penetrating comments. Finally, I would like to thank two anonymous referees for *Ethical Theory and Moral Practice*, whose excellent comments were of great help in improving the paper.

References

Ben-Moshe, N. (2020a). An Adam Smithian Account of Moral Reasons. *European Journal of Philosophy* 28(4): 1073-1087.

Ben-Moshe, N. (2020b). Hume's General Point of View: A Two-Stage Approach. *Pacific Philosophical Quarterly* 101(3): 431-453.

Ben-Moshe, N. (2020c). Making Sense of Smith on Sympathy and Approbation: Other-Oriented Sympathy as a Psychological and Normative Achievement. *British Journal for the History of Philosophy* 28(4): 735-755.

Ben-Moshe, N. (2021). Comprehensive or Political Liberalism? The Impartial Spectator and the Justification of Political Principles. *Utilitas*. Online First. https://doi.org/10.1017/S0953820820000394.

Brandt, R. B. (1955). The Definition of an "Ideal Observer" Theory in Ethics. *Philosophy and Phenomenological Research* 15(3): 407-413.

Brandt, R. B. (1959). *Ethical Theory: The Problems of Normative and Critical Ethics*. Engelwood Cliffs, NJ: Prentice-Hall, Inc.

Campbell, T. D. (1971). *Adam Smith's Science of Morals*. Oxon & New York: Routledge.

Carson, T. L. (1984). *The Status of Morality*. Dordrecht: D. Reidel Publishing Company (Philosophical Studies Series in Philosophy).

Darwall, S. (1998). Empathy, Sympathy, Care. *Philosophical Studies* 89(2-3): 261-282.

Enoch, D. (2005). Why Idealize? *Ethics* 115(4): 759-787.

Enoch, D. (2009). Can There Be a Global, Interesting, Coherent Constructivism about Practical Reason? *Philosophical Explorations* 12(3): 319-339.

Firth, R. (1952). Ethical Absolutism and the Ideal Observer. *Philosophy and Phenomenological Research* 12(3): 317-345.

Fleischacker, S. (2011). Adam Smith and Cultural Relativism. *Erasmus Journal for Philosophy and Economics* 4(2): 20-41.

Fleischacker, S. (2012). Sympathy in Hume and Smith: A Contrast, Critique, and Reconstruction. In C. Fricke & D. Follesdal (Eds.), *Intersubjectivity and Objectivity in Adam Smith and Edmund Husserl*. Heusenstamm, Germany: Ontos Verlag, pp. 273-311.

Fleischacker, S. (2013). Adam Smith's Moral and Political Philosophy. In E. N, Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. URL=<http://plato.stanford.edu/entries/smith-moral-political/>

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68(1): 5-20.

Griswold, C. L. (1999). *Adam Smith and the Virtues of Enlightenment*. New York: Cambridge University Press.

Johnston, M. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, Supplementary Volumes 63: 139-174.

Kawall, J. (2004). Moral Response-Dependence, Ideal Observers, and the Motive of Duty: Responding to Zangwill. *Erkenntnis* 60(3): 357-369.

Kawall, J. (2006). On the Moral Epistemology of Ideal Observer Theories. *Ethical Theory and Moral Practice* 9(3): 359-374.

Kornblith, H. (2012). *On Reflection*. Oxford: Oxford University Press.

Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society*, Supplementary Volumes 63: 113-137.

McDowell, J. (1998). Values and Secondary Qualities. In *Mind, Value & Reality*. Cambridge, MA & London: Harvard University Press, pp. 131-150.

Newton, J. (1788). *Thoughts upon the African Slave Trade*. URL=<https://cowperandnewtonmuseum.org.uk/wp-content/uploads/2020/07/thoughts-upon-african-slave-trade-john-newton.pdf >

Raphael, D. D. (2007). *The Impartial Spectator: Adam Smith's Moral Philosophy*. New York: Oxford University Press.

Rawls, J. (1999). *A Theory of Justice* (Revised Edition). Cambridge MA: Harvard University Press.

Sayre-McCord, G. (1994). On Why Hume's 'General Point of View' Isn't Ideal – And Shouldn't Be. *Social Philosophy and Policy* 11(1): 202-228.

Sayre-McCord, G. (2010). Sentiments and Spectators: Adam Smith's Theory of Moral Judgment. In V. Brown & S. Fleischacker (Eds.), *The Philosophy of Adam Smith: The Adam Smith Review* (Vol. 5). London & New York: Routledge, pp. 124-144.

Sayre-McCord, G. (2013). Hume and Smith on Sympathy, Approbation and Moral Judgment. *Social Philosophy and Policy* 30(1-2): 208-236.

Schliesser, E. (2017). *Adam Smith: Systematic Philosopher and Public Thinker*. New York: Oxford University Press.

Shafer-Landau, R. (2003). *Moral Realism: A Defence*. New York: Oxford University Press.

Smith, A. (1976). *The Theory of Moral Sentiments*. D. D. Raphael & A. L. Macfie (Eds.). Indianapolis: Liberty Fund.

Sobel, D. (2009). Subjectivism and Idealization. *Ethics* 119(2): 336-352.

Taliaferro, C. (1988). Relativizing the Ideal Observer Theory. *Philosophy and Phenomenological Research* 49(1): 123-138.

Vallentyne, P. (1996). Response-Dependence, Rigidification and Objectivity. *Erkenntnis* 44(1): 101-112.

Wiggins, D. (1998). A Sensible Subjectivism? In *Needs, Values, Truth: Essays in the Philosophy of Value* (3rd ed.). New York: Oxford University Press, pp. 185-214.

Williams, B. (1981). Internal and External Reasons. In *Moral Luck: Philosophical Papers 1973-1980*. Cambridge, UK: Cambridge University Press, pp. 101–113.

Williams, B. (1995). Internal Reasons and the Obscurity of Blame. In *Making Sense of Humanity and Other Philosophical Papers 1982-1993*. Cambridge, UK: Cambridge University Press, pp. 35–45.

Williams, B. (2001). Postscript: Some Further Notes on Internal and External Reasons. In E. Millgram (Ed.), *Varieties of Practical Reasoning*. Cambridge, MA: MIT Press, pp. 91–97.

Zangwill, N. (2003). Against Moral Response-Dependence. *Erkenntnis* 59(3): 285-290.