Ethic Theory Moral Prac (2015) 18:1059–1069
DOI 10.1007/s10677-015-9586-4

Towards a Cognitive Scientific Vindication of Moral Realism: The Semantic Argument

Abraham D. Graber[1]

Accepted: 11 March 2015 / Published online: 25 March 2015

Abstract In a methodological milieu characterized by efforts to bring the methods of philosophy closer to the methods of the sciences, one can find, with increasing regularity, meta-ethical arguments relying on scientific theory or data. The received view appears to be that, not only is it implausible to think that a scientific vindication of a non-mentalist moral semantics will be forthcoming but that evidence from a variety of sciences threatens to undermine non-mentalist views. My aim is to push back against this apparent consensus. The well-established phenomenon whereby moral judgments influence our attributions of putatively purely descriptive properties has come to be known as the Knobe Effect. Joshua Knobe has attempted to explain this surprising phenomenon by arguing that our folk psychological judgments are partially constituted by moral judgments. Drawing on an argument originally offered by Moore, I argue that if some instantiation of Knobe's explanatory strategy is accurate, we have good reason to believe that mentalist moral semantics are untenable.

Keywords The Knobe effect · The side-effect effect · Moral semantics · Non-mentalism · Moral realism

1 Introduction

With increasing regularity one can find meta-ethical arguments relying on scientific theory or data. The received view appears to be that, not only is it implausible to think that a scientific vindication of non-mentalism will be forthcoming (Shafer-Landau 2007), but that evidence from a variety of sciences threatens to undermine non-mentalist views (see, e.g., Levy 2006;

Abraham D. Graber is an Instructor of philosophy at Western Illinois University.

✉ Abraham D. Graber
agraber@gmail.com

[1] Department of Philosophy and Religious Studies, College of Liberal Arts and Sciences, Western Illinois University, 456 Morgan, Macomb, IL 61455, USA

Prinz 2007; Street 2006). My aim is to push back against this apparent consensus. I will briefly sketch the contours of an emerging research program in cognitive science then, drawing on one of Moore's many criticisms of subjectivism, argue that if the central claims of this research program are accurate, the dominant approach to anti-realist semantics is untenable.

2 Person as Scientist, Person as Moralist

2.1 The Explanadum

In "Intentional Action and Side Effects in Ordinary Language" Knobe demonstrated that a surprising relationship holds between moral judgments and attributions of intentionality. On the commonsense view, moral permissibility is partially determined by the intentions of an agent. Thus, if Diane did not intend to hurt a puppy but did so accidentally, we tend to think that Diane is less morally culpable than if she acted intentionally. Commonsensically, judgments about the moral permissibility are subsequent to judgments of intentionality.

Knobe's work suggests that the picture is not this neat. Consider the following vignette:

> The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
> The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'
> They started the new program. Sure enough, the environment was harmed. (Knobe 2003)

Test subjects were asked if the chairman of the board intentionally harmed the environment. By and large, subjects answered "yes."

Consider a modified version of the above vignette. Replace every instance of the phrase "harm the environment" (and cognates) with "help the environment" (and cognates). While maintaining her concern for profit and indifference towards the environment, the CEO approves an environmentally friendly program. By and large, subjects given this modified vignette judged that the chairman *did not* intend to *help* the environment.

The vignettes are structurally identical. The only difference is that in the first instance the chairman acted in a morally questionable way—harming the environment in order to increase profits—whereas in the latter case the chairman found herself in morally untroubled waters. Judgments about the moral permissibility of an action appear to influence judgments about intentionality. This phenomenon, whereby moral judgments influence the attribution of putatively descriptive properties, has come to be known as the *Knobe Effect*.

One might think that the Knobe Effect is isolated to attributions of intentionally. Research suggests otherwise. The effect has been observed with attribution of all of the following: deciding, desiring, being in favor of, advocating, causing, being free, knowing that, and more. (Knobe 2010; Beebe and Jensen 2012) The Knobe Effect is a "general tendency, whereby moral judgments impact the application of a whole range of different concepts used to pick out mental states and processes" (Knobe 2010).

Evidence as to the extent of the Knobe Effect is still forthcoming; however, my aim is to show that it would be very good news for the non-mentalist if the central tenets of Knobe's research program turned out to be true. Thus, alongside Knobe I will assume that the effect is ubiquitous, particularly with regard to the attribution of folk psychological predicates.

2.2 Motivational Bias and Conversational Pragmatics: Two Unpromising Explanatory Strategies

Large swaths of putatively descriptive judgments are influenced by moral considerations. An explanation is called for. There are, broadly speaking, two explanatory strategies one might take. One might think that our fundamental competencies are purely descriptive and that moral considerations lead to performance errors. Alternatively, one might think that the Knobe Effect demonstrates something surprising about our core competencies: normative considerations play a role in the deep structure of our competence with folk psychological predicates. The latter explanatory strategy is radically revisionary; we would do well to consider the former. Two primary explanations have been offered to demonstrate that the Knobe Effect tracks performance errors.

The first explanation holds that the Knobe effect is a consequence of *motivational bias*: "once strong negative reactions have been evoked, people view the relevant evidence in a way that justifies their desire to blame the source of those reactions" (Alicke 2008). Subjects come to dislike the chairman of the board and try to justify their negative affect by forming the belief that the chairman of the board intended to harm the environment. Similar explanations of the Knobe effect have been offered by (Malle 2006; Malle and Nelson 2003; Nadelhoffer 2006). All of these views draw on the idea that the Knobe effect is a consequence of people attempting to cognitively justify their conative responses. The general phenomenon whereby people attempt to justify their conative responses, falling under the heading of *motivated reasoning*, has been well established (Kunda 1990).

Though the *motivational bias* hypothesis offers an elegant explanation of the Knobe Effect, the empirical evidence does not support the hypothesis. Young et al. replicated Knobe-style experiments with subjects who, as a result of brain damage, are incapable of normal affective responses. The Knobe Effect was still present; a result we would not expect if the Knobe Effect were the consequence of an attempt to justify negative affect (Young et al. 2006).

Further, were the motivational bias hypothesis correct, we would expect the Knobe Effect to disappear when a dislikable agent brought about positive outcomes. The Knobe Effect has been observed in attributions of causation. Subjects are more likely to judge that an agent who acted in a morally questionable way caused a positive outcome than they are to judge that an agent who acted in a morally neutral way caused the same positive outcome (Hitchcock and Knobe 2009). If the Knobe Effect is a consequence of our aiming to justify negative affect, we should not be more likely to credit a dislikeable agent with bringing about something good.

The second explanatory strategy holds that the Knobe Effect is a consequence of *conversational pragmatics*. Subjects import considerations, not about what the questions and answers *mean*, but about how a certain answer to a given question would be likely to be *understood*: "[Subjects] know that their blame is stronger and more effective at discouraging [morally reprehensible actions], if the chairman is said to have done the action *intentionally* (Adams and Steadman 2004)." Like the motivational bias hypothesis, the *conversational pragmatics hypothesis* draws on a relatively well-understood phenomenon in order to explain novel observations; however, also like the motivational bias hypothesis, the conversational pragmatics hypothesis is unsupported by the empirical evidence.

Zalla, Machery, and Leboyer replicated Knobe-style experiments with subjects on the autism spectrum (Zalla, Machery, and Leboyer 2008). Individuals on the autism spectrum are well known to have difficulties with conversational pragmatics, "tending to answer questions in the most literal possible way" (De Villiers, Stainton, and Szatmari 2006). Were the Knobe Effect an artifact of conversational pragmatics, we would not expect to observe it in

these subjects. Contra the conversational pragmatics hypothesis, the Knobe Effect is present in individuals on the autism spectrum (Knobe 2010).

There are a huge number of potentially plausible explanations for the Knobe Effect whereby the effect is nothing more than a performance error. Debunking all such defusing explanations is a task of overwhelming magnitude. Nonetheless, things do not look good for the primary explanations whereby the Knobe Effect is a mere performance error. This suggests that we take a different approach to explaining the Knobe Effect:

> Instead of focusing on the interfering factors, we will try looking at the competence itself. The aim will be to show that something about the very nature of … [our] competence [with folk psychological terms] is allowing people's moral judgments to influence their intuitions. (Knobe 2010)

2.3 The Knobe Effect and Core Competencies

If Knobe wishes to show that the Knobe Effect demonstrates something about our core capacities, he will have to offer an explanation where the mechanism responsible for the Knobe Effect is central to our competency with folk psychological terms.

(Knobe continues to offer increasingly refined version of the explanatory strategy sketched below. The most detailed, nuanced, and plausible of these explanations would require a lengthy summary and remains unpublished (Szabo and Knobe forthcoming). For these reasons, I have chosen to focus on some of Knobe's older work. Importantly, central features of Knobe's explanatory strategy remain constant. I highlight these in the following discussion.)

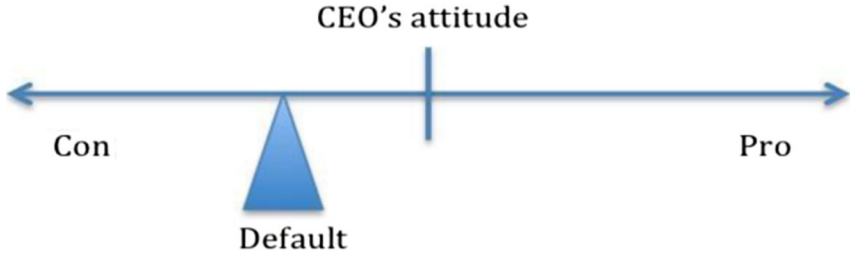
Knobe structures his explanation around the fact that modal judgments play a central role in the way we understand the world:

> [W]e make sense of the things that actually happen by considering *other ways things might have been*… Our ability to pick out just certain specific alternatives and ignore others is widely regarded as a deeply important aspect of human cognition, which shapes our whole way of understanding the objects we observe… A number of studies have shown that people's selection of alternative possibilities can be influenced by their *moral judgments*… Because people's moral judgments influence the selection of alternative possibilities, these moral judgments end up having a pervasive impact on the way people make sense of human beings and their actions. (Knobe 2010)

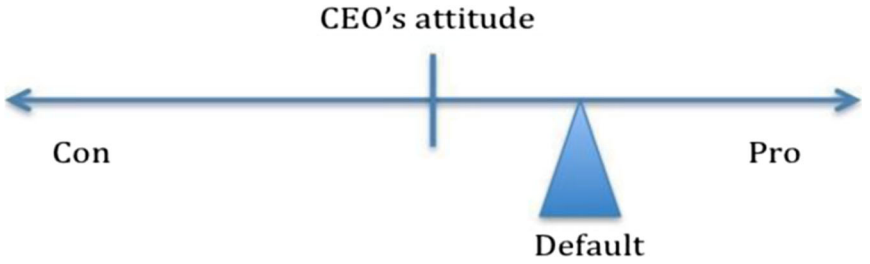
Consider a revised version of the vignette involving the chairman of the board and potential environmental harm. In this vignette, instead of asking if the CEO *intended* to harm/help the environment, we ask if the CEO was *in favor of* harming/helping the environment. Knobe suggests that we understand attributions of *being in favor of* not in terms of "a simple dichotomy" between *being in favor of* and not *being in favor of*, but rather "in terms of a whole *continuum* of different attitudes an agent might hold" (Knobe 2010). This continuum can be depicted as a line running from con-attitudes (directed at some state-of-affairs) on the left to pro-attitudes (directed at the same state-of-affairs) on the right.

Somewhere on the continuum from *con* attitudes to *pro* attitudes there must be a threshold that marks the line between *being in favor of such-and-such* and ***not*** *being in favor of such-and-such.* Knobe labels this threshold the *default*. The placement of the default is not static. Rather, "people's moral judgments affect their intuitions *by shifting the position of the default*" (Knobe 2010). We are now in a position to explain the Knobe effect.

Intending to harm the environment is morally closed; attitudes at the far *pro* end of the spectrum are not considered among the relevant possibilities. This pushes the default for *being in favor of* ***harming*** *the environment* towards the *con* end of the spectrum. Thus, mere indifference towards harming the environment falls on the *pro* side of the default and consequently counts as *being in favor of* ***harming*** *the environment.*



When we ask if the chairman *was in favor of* ***helping*** *the environment* an inverted style of the same reasoning works. Certain kinds of *con*-attitudes towards helping the environment are morally closed, pushing the default towards the *pro* end of the spectrum. Thus, mere indifference towards helping the environment falls on the *con* side of the default and fails to count as *being in favor of* ***helping*** *the environment.*



As noted above, this summary does not capture Knobe's most recent work. Nonetheless, all of Knobe's attempts to explain the Knobe Effect share core features. Of particular importance for us, all of Knobe's attempts to explain the Knobe Effect offer an explanation whereby our folk psychological judgments (e.g. the judgment that *the CEO intended to hurt the environment*) are partially constituted by modal judgments (included amongst these modal judgments are moral judgments). It is perhaps best to understand the preceding discussion as offering a (very) broad strokes sketch of Knobe's explanatory strategy. There are a variety of ways in which Knobe can fill out the details and we can expect the details to change as Knobe's research program matures.

If accurate, this explanatory strategy calls for a radical revision in our understanding of the psychological processes that underlie our property attributions. Knobe is optimistic that the Knobe Effect is cognitively ubiquitous, particularly amongst attribution of folk psychological predicates. If the empirical evidence bears out Knobe's optimism, we may be in a position to reduce an enormous range of linguistic competencies to a mere handful of modal concepts with a few bells and whistles attached. At the foundation of this sparse competency structure, one will find moral concepts, for "[e]ven the [psychological] processes that look most "scientific" actually take moral considerations into account. It… [would seem] that we are moralizing creatures through and through" (Knobe 2010).

3 The Prospects for a Cognitive Scientific Vindication of a Non-mentalist Moral Semantics

I have now offered a rough sketch of an emerging research program in cognitive science. Central to this research program is the claim that our folk psychological judgments are partially constituted by moral judgments. In the remainder of this paper I will consider the upshot for the debate over the correct account of moral semantics. Towards this end I will assume (1) that the best explanation of the Knobe Effect will be some instantiation of the explanatory strategy sketched above and (2) that Knobe is correct that the Knobe Effect is ubiquitous amongst our folk psychology judgments. Each assumption represents a pillar of Knobe's emerging research program; acceptance of these two assumptions constitutes taking this emerging research program seriously.

3.1 Mentalist and Non-mentalist Moral Semantics

In the remainder of this paper I will argue that, if Knobe is correct, mentalist approaches to moral semantics are untenable. The mentalist family is composed of those meta-ethical views that hold that moral predicates can be analyzed, without moral remainder, in terms of folk psychological states. Mentalism is a wide and varied family of views the members of which compose the bulk of anti-realist approaches to moral semantics. Each of the following is a member of the mentalist family. Gibbard's expressivism is a version of mentalism. Gibbard aims to reduce moral thought to expressions of mental states, "[T]o think that compassion is good is to *accept a norm* that says to desire compassion," where acceptance should be understood as a certain kind of folk psychological state (Gibbard 2009: 7). Blackburn's expressivism is also a version of mentalism: "to think that x is wrong is to disapprove of x and to disapprove of those who fail to share this disapproval" (Sinclair 2009; Blackburn 1998, 2006). Expressivist views are far from the only members of the mentalist family. Various versions of cognitivism also fit the mold. Consider a view recently defended by Prinz: "When I say something is wrong, I refer (perhaps unwittingly) to the property of causing emotions of blame in me" (Prinz 2006: 35). Like Gibbard and Blackburn, Prinz aims to analyze moral predicates, without remainder, in terms of folk psychological states. While Prinz is a senti-mentalist, constructivist views also fall within the broad purview of mentalism: "metaethical constructivism asserts a counterfactual dependence of value on the *attitudes* of valuing creatures; it understands reason-giving status as conferred upon things by us" (Street 2010: 371, emphasis added).

Non-mentalist approaches to moral semantics deny the mentalist's claim. Thus, the non-mentalist holds that moral predicates cannot be analyzed, without normative remainder, in terms of folk psychological states. Any attempt to analyze a moral predicate in terms of folk psychological states will either fail or include a normative term in the analysans.

The relationship between moral realism and mentalism is not straightforward. Some versions of realism are members of the mentalist family. Thus, in describing his version of moral realism (defended in [Railton 1986]) Railton writes: "the ultimate ground of normativity [is located] in the *affective dispositions* of agents" (Darwall, Gibbard, and Railton 1992: 176, emphasis added). Similarly, the view Michael Smith defends in *The Moral Problem* (1994) appears to be a realist version of mentalism.

Furthermore, there are anti-realist meta-ethical views that fall in the non-mentalist camp. Loeb has defended a view he labels "incoherentism." On this view, the correct account of moral semantics is that there is no correct analysis of moral semantics. In virtue of the variety

of incompatible ways in which moral language is used, moral language is fundamentally incoherent (Loeb 2008). Loeb's view is clearly a version of non-mentalism. Loeb does not believe that moral predicates can be analysed, without normative remainder, in terms of folk psychological states. Loeb doesn't think that moral predicates can be analyzed at all!

(In an unpublished paper, Gilbert Harman (2012) defends a version of moral relativism which he claims "is not a linguistic or conceptual thesis" (15). Depending on how we understand the view he defends in that paper, it may also qualify as a non-mentalist version of moral anti-realism.)

Though realism and non-mentalism are orthogonal, it would nonetheless be a significant victory for the moral realist were non-mentalism to be vindicated. The most well known versions of moral realism all endorse a non-mentalist semantics (Brink 1989; Boyd 1988; Enoch 2011; Shafer-Landau 2003) while the primary versions of anti-realism (non-cognitivism, cognitivist sentimentalism, and constructivism) are all members of the mentalist family. My aim in this paper is to consider what implications an explanation of the Knobe effect would have for the plausibility of the mentalist's semantic program. Insofar as the primary anti-realist accounts of moral semantics are members of the mentalist family and the primary realist accounts of moral semantics are members of the non-mentalist family, the arguments I present here will have important implications for the debate between the moral realist and the moral anti-realist.

3.2 Moore and an Untenable Approach to Moral Semantics

While Moore is best known for his open question argument, it is only one of his many contributions to the non-mentalist's cause. In this portion of the paper I will briefly review one of Moore's arguments against a specific version of subjectivism; the remainder of the paper is dedicated to demonstrating that, assuming Knobe is correct about the best explanation of the Knobe Effect, Moore's argument applies to a wide range of mentalist semantic accounts.

Moore distinguishes between (at least) two versions of subjectivism. The first is comparatively more familiar from the contemporary literature: "it may be held that whenever any man [sic] asserts an action to be right or wrong, what he is asserting is merely that he *himself* has some particular feeling towards the action in question" (Moore 2014:39). Contemporary versions of this meta-ethical view are defended by Prinz (2006, 2007), amongst others. To the best of my knowledge, the other variety of subjectivism Moore considers has no contemporary defenders (though one frequently hears undergraduates suggest some version of this view). This view holds that "when we judge an action to be right or wrong what we are asserting is merely that somebody or other *thinks* it to be right or wrong" (Moore 2014:54).

It is no accident that this latter view lacks defenders. It is fatally flawed for "it is in all cases totally impossible that, when we believe a given thing, *what* we believe should merely be that we (or anybody else) have the belief in question… because, if it were the case, we should not be believing anything at all." (Moore 56). Suppose we think that the statement *killing is wrong* as said by Sally just means *Sally believes that killing is wrong*. Notice that the analysans contains the analysandum, "killing is wrong." Thus, we can replace "killing is wrong" in *Sally believes that killing is wrong* with its analysans, netting us: *Sally believes that Sally believes that killing is wrong*. But the analysans again re-occurs in the analysandum! "[W]hat I am believing will turn out to be that somebody believes, that somebody believes, that somebody believes, that somebody believes… *an infinitum*" (ibid.). The consequence is that "I shall never get to anything whatever which is *what* is believed" (ibid.).

It may be difficult to see how Moore's criticism could generalize to more *prima facie* plausible mentalist semantics. In the remainder of the paper I will argue that, if we grant Knobe's explanation of the Knobe Effect, Moore's argument is more far reaching than it at first appears.

3.3 Expanding Moore's Argument

The argument for this conclusion is surprisingly simple. On Knobe's picture, making a folk psychological judgment should be understood as judging that an attitude falls on one side (or the other) of the default. But the position of the default is partially determined by moral considerations. Thus, if Knobe is correct, judgments about folk psychological states are partially constituted by moral judgments.

By way of illustration, consider how an attempted (naïve) analysis of "right" might go: the statement *killing is wrong*, said by a speaker A, is true just in case A disapproves of killing. So far, there is no problem. Nothing Knobe says rules out the veracity of such an analysis. A problem does arise, however, if one hopes to analyze *away* moral properties. Put another way, a problem arises if the proponent of the above naïve analysis of "right" wants to offer an analysis in which moral predicates do not appear in the analysans.

If Knobe is correct, we can now further analyze "A disapproves of killing." A disapproves of killing only if A's attitude towards killing falls on the appropriate side of the default. If we want to analyze the statement *killing is wrong*, as said by A, we need to know what it means for *A's attitude towards killing to fall on the appropriate side of the default*. A's attitude towards killing falls on the appropriate side of the default just in case (a) A's attitude towards killing has a certain valence and strength and (b) A's statement is made in a context where the relevant modal judgments place the default in such-and-such a position. Importantly, the modal judgments mentioned in (b) include *moral judgments*. Thus, the analysis of *A's disapproving of killing* includes a claim about moral judgments.

If Knobe is correct, every time we attempt to analyze the meaning of a moral judgment in terms of folk psychological states, moral predicates will re-occur in the analysans. *Either* our rock bottom analysis of A's claim that *killing is wrong* will include moral predicates *or* it will continue ad infinitum. If our analysis includes moral predicates, we have failed to analyze, without moral remainder, moral predicates in terms of folk psychological states. Alternatively, if our analysis continues indefinitely, when we form moral beliefs, we are "not be believing anything at all" (Moore 56) for the content of the belief is, in principle, unspecifiable.

It is important to see that this problem is not unique to the naïve version of subjectivism offered above. If Knobe is correct, a similar story can be told about nearly every folk psychological predicate. No matter how nuanced one's moral semantics, if one's analysans contains folk psychological terms, one's analysans contains a disguised reference to moral properties. No attempt to reduce morality to mental states can be successful.

The careful reader likely arched an eyebrow upon reading the preceding paragraph; if Knobe can only tell a similar story about *nearly every* folk psychological predicate, does that mean, even on his view, there are folk psychological states that can be analyzed without reference to moral properties? Even Knobe's radical account cannot do without unanalyzable folk psychological concepts. Knobe's explanation of the Knobe Effect as observed with attributions of "being in favor of…" required positing a continuum between *pro* and *con* attitudes. These are clearly folk psychological concepts, so some folk psychological judgments must not be analyzable in terms of moral judgments. If this is the case, it may seem that mentalist semantics

are alive and well. Instead of attempting to analyze moral predicates in terms of approval, disapproval, assent, acceptance, etc., the mentalist need merely analyze moral judgments in terms of whatever folk psychological states escape Knobe's reductive program.

While the mentalist is correct that some folk psychological states escape Knobe's reduction, it is unlikely to help her case. Mentalist moral semantics require comparatively thick folk psychological notions. I can *approve* of going to the gym, or *disapprove* of eating donuts. Similarly, I can feel a *pro-attitude* towards, e.g., pencils and a *con-attitude* towards, e.g. pens. The mentalist cannot let just *any* con-attitude be the truth-maker of moral claims; otherwise, one gets extremely implausible moral consequences, e.g. the impermissibility of eating donuts and the impermissibility of writing with a pen. The types of folk psychological ascriptions that Knobe's explanatory strategy requires are not thick enough to ground a plausible mentalist semantics.

4 Moral Mentalism vs. Normative Mentalism

So far I have assumed that specifically *moral* considerations drive the Knobe effect. There is, however, some reason to doubt that this is the case. Several authors have argued that broadly normative considerations, of which moral norms are only a subset, are at the heart of the Knobe effect (see, e.g., [Holton 2010; Guglielmo 2010; Robinson, Stey, and Alfano forthcoming]). Suppose some such account is accurate. What implications does this have for my argument for non-mentalism?

Much will depend on the specifics of the view in question. So long as moral norms are included amongst the norms that determine the position of the default, the argument for non-mentalism I have developed is perfectly compatible with the view that the Knobe effect is a consequence of norms generally, and not moral norms specifically. So long as moral judgments play a role in determining the position of the default when we make judgments about folk psychology, moral judgments are conceptually prior to folk psychological judgments. Consequently, no attempt to reduce moral judgments to folk psychological judgments will be successful. This will remain the case even if various non-moral normative judgments also play an important role in positioning the default.

There are, however, versions of the view that are potentially problematic for the argument for non-mentalism I have advanced. One might think–though I know of nobody who has advanced this view–that though normative judgments determine the position of the default for all folk psychological judgments, moral judgments are only relevant to the position of the default in a subset of cases. Such a view would allow the mentalist to construct a moral semantics grounded in those folk psychological judgments that are not partially constituted by moral judgments.

Nonetheless, the non-mentalist has good reason to hope for the truth of even a version of Knobe's view restricted in the above way. These days the discipline of metaethics is in the business of studying norms generally, not merely moral norms (Scanlon 2014). Thus, we can distinguish between two versions of mentalism: moral mentalism and normative mentalism. Moral mentalism is a subset of normative mentalism. (Similarly, moral non-mentalism is a subset of normative non-mentalism.) The moral mentalist holds that *moral* terms can be analyzed, without normative remainder, in terms of folk psychological states. The normative mentalist holds that *normative* terms can be analyzed, without normative remainder, in terms of folk psychological states.

If a restricted version of Knobe's view is correct then moral mentalism may be true; however, normative mentalism will still be false. If normative judgments play an ineliminable role in our application of folk psychological terms, there is no hope of analyzing normativity in terms of folk psychology. The significance of establishing the truth of normative non-mentalism should not be understated. Establishing the truth of a non-mentalist normative semantics would be an accomplishment only slightly less momentous than establishing the truth of a non-mentalist moral semantics.

5 Conclusion

I have now argued that, if Knobe is correct about the best explanation of the Knobe Effect, moral mentalism faces a devastating problem: moral predicates cannot be analyzed, without moral remainder, in terms of folk psychological predicates. Knobe's research program is, however, still young. Furthermore, there are a variety of ways a mentalist could respond to the arguments I have offered. Rather than offering a knockdown argument against mentalism, I hope to be taking the first steps in the development of a research program that will, eventually, operationalize a certain sort of robust moral realism. Drawing on Knobe's work, I have offered a proof of concept of an empirical defense of the semantic program of a version of non-mentalism. Once we have good reason to think that it is possible to provide such a defense, it only seems appropriate for the non-mentalist to make the operationalization of non-mentalism a centerpiece of her meta-ethical project.

Acknowledgments Many thanks are owed to Richard Fumerton for his help developing the arguments in this paper and to Joshua Knobe for his encouragement at the early stages of this project. Above all, I owe thanks to Jessica Schwartz for her unfailing support.

References

Adams F, Steadman A (2004) Intentional action in ordinary language: core concept or pragmatic understanding? Analysis 64:173–181
Alicke MD (2008) Blaming badly. J Cogn Cult 8:179–186
Beebe J, Jensen M (2012) Surprising connections between knowledge and action: The robustness of the epistemic side-effect effect. Philos Psychol 25:689–715
Blackburn S (1998) Ruling passions. Clarendon, Oxford
Blackburn S (2006) Antirealist expressivism and quasi-realism. In: Copp D (ed) The Oxford handbook of ethical theory. Oxford University Press, Oxford
Boyd R (1988) How to be a moral realist. In: Sayre-McCord G (ed) Essays on moral realism. Cornell University Press, Ithaca, pp 181–228
Brink D (1989) Moral realism and the foundations of ethics. Cambridge University Press, Cambridge
Darwall S, Gibbard A, Railton P (1992) Toward fin de siecle ethics: some trends. Philos Rev 101:115–190
De Villiers J, Stainton R, Szatmari P (2006) Pragmatic abilities in autism spectrum disorder: a case study in philosophy and the empirical. Midwest Stud Philos 31:292–317
Enoch D (2011) Taking morality seriously: a defense of robust realism. Oxford University Press, Oxford
Gibbard A (2009) Thinking how to live. Harvard University Press, Cambridge
Guglielmo S (2010) Questioning the influence of moral judgment. Behav Brain Sci 33:338–339
Harman G (2012). Moral realism is moral relativism. http://www.princeton.edu/~harman/Papers/Relativism_Realism.pdf.
Hitchcock C, Knobe J (2009) Cause and norm. J Philos 106:587–612
Holton R (2010) Norms and the Knobe effect. Analysis 70:417–424

Knobe J (2003) Intentional action and side effects in ordinary language. Analysis 63:190–194
Knobe J (2010) Person as scientist, person as moralist. Behav Brain Sci 33:315–365
Kunda Z (1990) The case for motivated reasoning. Psychol Bull 108:480–498
Levy N (2006) Cognitive scientific challenges to morality. Philos Psychol 19:567–587
Loeb D (2008) How to pull a metaphysical rabbit out of a semantic hat. In: Walter S-A (ed) Moral psychology, vol 2. MIT Press, Cambridge, pp 355–385
Malle B (2006) Intentionality, morality, and their relationship in human judgment. J Cogn Cult 6:87–112
Malle B, Nelson S (2003) Judging mens rea: the tension between folk concepts and legal concepts of intentionality. Behav Sci Law 21:563–580
Moore GE (2014) Ethics. Createspace Independent Publishing Platform, Lexington, KY
Nadelhoffer T (2006) Bad acts, blameworthy agents, and intentional actions: some problems for jury impartiality. Philos Explor 9:203–220
Prinz J (2006) The emotional basis of moral judgments. Philos Explor 9:29–43
Prinz J (2007) The emotional construction of morals. Oxford University Press, Oxford
Railton P (1986) Moral realism. Philos Rev 95:163–207
Robinson B, Stey P, Alfano M (Forthcoming). Reversing the side-effect effect: the power of salient norms. Philos Stud
Scanlon TM (2014) Being realistic about reasons. Oxford University Press, Oxford
Shafer-Landau R (2003) Moral realism: a defence. Clarendon, Oxford
Shafer-Landau R (2007) Moral and theological realism: the explanatory argument. J Moral Philos 4:311–329
Sinclair N (2009) Recent work in expressivism. Analysis 69:136–147
Smith M (1994) The moral problem. Blackwell Publishing Ltd, Malden, MA
Street S (2006) A Darwinian dilemma for realist theories of value. Philos Stud 127:109–166
Street S (2010) What is constructivism in ethics and metaethics? Philos Compass 5:363–384
Szabo ZG, Knobe J (Forthcoming). Modals with a taste of the deontic. Semant Pragmat
Young L, Cushman F, Adolphs R, Tranel D, Hauser M (2006) Does emotion mediate the effect of an action's moral status on its intentional status? Neuropsychological evidence. J Cogn Cult 6:291–304
Zalla T, Machery E, Leboyer M (2008). Intentional action and moral judgment in Asperger Syndrome and high functioning Autism. Paper presented at the Winter Workshop 2008 on Games, Experiments and Philosophy, Jena, Germany