Discussion Note

Concept Formation in Ethical Theories: Dealing with Polar Predicates

by Sebastian Lutz

Journal of Ethics & Social Philosophy

Discussion Note | August 2010
URL: www.jesp.org

Concept Formation in Ethical Theories: Dealing with Polar Predicates

Sebastian Lutz

IN "A DANGER OF DEFINITION: Polar Predicates in Metaethics,"[1] Mark Alfano (2009) concludes that the response-dependence theory of Prinz and others and the fitting-attitudes theory first articulated by Brentano are false because they imply empirically false statements. He further concludes that these statements cannot be avoided by revising the definitions of the terms "good" and "bad" used in the two theories. In this note, I strengthen Alfano's first conclusion by arguing that the two theories are false even if they imply empirically true but conceptually contingent statements, and show how, contrary to his second conclusion, the theories can avoid both empirically false and conceptually contingent implications.

1. The Case Against Response-dependence and Fitting-attitudes Theories

1.1 An empirical inconsistency

According to Alfano, response-dependence and fitting-attitudes theories contain explicit definitions for "good" and "bad." In response-dependence theory, something is good (bad) if and only if someone is disposed to have a positive (negative) sentiment toward it upon careful reflection (p. 3); in fitting-attitudes theory, something is good (bad) if and only if it would be fitting to take an approbative (disapprobative) attitude toward it (p. 8). Since "good" and "bad" are polar predicates, they are contraries; that is, nothing is both good and bad.

The two theories share a structure with all theories that contain explicit definitions of contraries, and which can be expressed as the simple theory T (p. 1):

$$(1) \qquad \forall x[Fx \leftrightarrow \varphi(x)]$$
$$(2) \qquad \forall x[\text{un-}Fx \leftrightarrow \psi(x)]$$
$$(3) \qquad \neg\exists x(Fx \wedge \text{un-}Fx)\ .$$

Postulates (1) and (2) are the explicit definitions of "good" (F) and "bad" (un-F). Postulate (3) is the claim that the two defined terms are contraries. If the definientia of these terms are not themselves contraries, then the world can turn out to be such that they are co-instantiated. At the core of Alfano's article (pp. 4-7, 9f) lies an argument for their co-instantiation:

$$(4) \qquad \exists x[\varphi(x) \wedge \psi(x)]\ .$$

[1] Unless otherwise indicated, page and section numbers refer to this article.

This claim is inconsistent with *T*; therefore, Alfano concludes, *T* is false.

Alfano further argues against six possible responses to this inconsistency:

(I) Dialetheism is too high a price to pay (p. 2).

(II) Giving up postulate (3) means that some things are both good and bad, which is not better than outright dialetheism (p. 7).

(III) Giving up postulate (1) or (2) results in an ethical theory that has nothing to say about either "good" or "bad" (p. 7).

(IV) Changing the definitions to make the empirical claim (4) false leaves the ethical theory, at least in principle, vulnerable to empirical refutation (p. 2).

(V) One of the two defined terms could be made into a trouser-word by introducing the new definition

(5) $\forall x\{Fx \leftrightarrow [\varphi(x) \wedge \neg\text{un-}Fx]\}$

or

(6) $\forall x\{\text{un-}Fx \leftrightarrow [\psi(x) \wedge \neg Fx]\}$,

but the choice between them is arbitrary and therefore ad hoc: There is no plausible argument for "good" being prior to "bad," and vice versa (p. 8).

(VI) Changing one of the two definitions to yield contradictories, that is, changing postulate (1) into $\forall x(Fx \leftrightarrow \neg\text{un-}Fx)$ or postulate (2) into $\forall x(\text{un-}Fx \leftrightarrow \neg Fx)$, is less plausible than the introduction of a trouser-word (p. 7f).

Thus, Alfano concludes that neither of the two theories can be saved.

1.2 A conceptual inconsistency

Alfano's rebuttal IV suggests a way of strengthening his argument that the two theories are false. Specifically, it is not necessary to establish that (4) is true, only that its contradictory,

(7) $\neg\exists x[\varphi(x) \wedge \psi(x)]$,

is conceptually contingent. It may be only an empirical (but not a conceptual) truth, for example, that the disposition to have a positive sentiment is contrary to a disposition to have a negative sentiment. Similarly, it might not be a conceptual truth that a fitting approbative attitude is contrary to a fitting disapprobative attitude.[2] So while definitions (1) and (2) ensure that (7)

[2] Note that Alfano (§3) argues that it is sometimes fitting to have both an approbative and a disapprobative attitude, even though for his argument, he only needs to establish that it is

entails (3), one could argue analogously to a consideration by Wlodek Rabinowicz (2008, p. 40) that these definitions are not satisfactory as complete reductions of F and un-F because they would reduce the conceptual truth (3) to the non-conceptual truth (7).[3]

But the problem is more severe. For assume that (7) is not conceptually true. Response-dependence and fitting-attitudes theories are meant to capture the concepts "good" and "bad"; in other words, definitions (1) and (2) are conceptually true. Since postulate (3) is conceptually true and, in connection with (1) and (2), entails (7), (7) is conceptually true as well, which contradicts the assumption.[4]

Alfano argues that response-dependence and fitting-attitudes theories are false beyond repair by identifying an empirical implication of T, arguing that this implication is false, and arguing that the postulates for F and un-F cannot be changed to avoid empirical implications. Under this abstract description of his argument, I can apply suggestions for the formation of concepts that were developed in the philosophy of science, notably by Rudolf Carnap and Marian Przełęcki, in order to arrive at new postulates that are consistent with (4) and avoid all of Alfano's objections to responses I–VI. That is, I aim to show that the postulates for F and un-F can be changed to avoid empirical and conceptual inconsistencies, and that therefore these theories can be salvaged.

2. Isolating the Empirical Content of Ethical Theories

The theory T can be equivalently reformulated as one necessary and one sufficient condition for F,

(8a) $\forall x[\varphi(x) \rightarrow Fx]$
(8b) $\forall x\{[\neg\varphi(x) \vee \psi(x)] \rightarrow \neg Fx\}$,

and one necessary and one sufficient condition for un-F:

(9a) $\forall x[\psi(x) \rightarrow \text{un-}Fx]$

sometimes both fitting to have an approbative attitude and fitting to have a disapprobative attitude. The latter claim is the correct paraphrase of Alfano's formula (20) (p. 9); the former is Alfano's paraphrase.

[3] I thank an anonymous referee for pointing out this problem and its discussion by Rabinowicz.

[4] Since all and only conceptual truths are conceptually necessary, the argument can be expressed (using "$\Box$" for "it is conceptually necessary that") as follows: $\{\Box\forall x[Fx \leftrightarrow \varphi(x)], \Box\forall x[\text{un-}Fx \leftrightarrow \psi(x)], \Box\neg\exists x(Fx \wedge \text{un-}Fx)\} \vDash \Box\neg\exists x[\varphi(x) \wedge \psi(x)]$, which contradicts $\neg\Box\neg\exists x[\varphi(x) \wedge \psi(x)]$, i.e., the assumption that (7) is not conceptually necessary. Since for conceptual necessity $\exists x[\varphi(x) \wedge \psi(x)] \vDash \neg\Box\neg\exists x[\varphi(x) \wedge \psi(x)]$ and $\neg\Box\neg\exists x[\varphi(x) \wedge \psi(x)] \nvDash \exists x[\varphi(x) \wedge \psi(x)]$ hold, the assumption is weaker than what Alfano needs to establish for his argument.

(9b) $\forall x\{[\neg\psi(x) \vee \varphi(x)] \rightarrow \neg\text{un-}Fx\}$.

Carnap (1936, §8) discusses postulates of the kind

(10a) $\forall x[\xi(x) \rightarrow Qx]$
(10b) $\forall x[\chi(x) \rightarrow \neg Qx]$

as "reduction pairs" and notes that together they imply the "representative sentence"

(11) $\forall x\neg[\xi(x) \wedge \chi(x)]$.

This sentence does not contain Q if, as is assumed, neither ξ nor χ does (Carnap 1936, §10, p. 451). As Alfano's rebuttal V shows, he assumes that neither φ nor ψ contain F or un-F; the conditions (8) and (9) are therefore reduction pairs. Both reduction pairs' representative sentence (7) is just the contradictory of Alfano's empirical claim (4).

Przełęcki (1961, p. 136) suggests the replacement of the reduction pair (10) by

(12a) $\forall x\{[\xi(x) \wedge \neg\chi(x)] \rightarrow Qx\}$
(12b) $\forall x\{[\chi(x) \wedge \neg\xi(x)] \rightarrow \neg Qx\}$

for two reasons. First, together with the representative sentence (11), these sentences are equivalent to the original reduction pair (10). Second, their own representative sentence is a tautology. This second point is important because reduction pairs that fulfill this condition are conservative as defined in the theory of definition, and so they do not have implications for terms other than Q (see, for example, Belnap 1993).

Applied to the two reduction pairs (8) and (9), Przełęcki's suggestion yields the new reduction pairs

(13a) $\forall x\{[\varphi(x) \wedge \neg\psi(x)] \rightarrow Fx\}$
(13b) $\forall x[\neg\varphi(x) \rightarrow \neg Fx]$

and

(14a) $\forall x\{[\psi(x) \wedge \neg\varphi(x)] \rightarrow \text{un-}Fx\}$
(14b) $\forall x[\neg\psi(x) \rightarrow \neg\text{un-}Fx]$.

By design of Przełęcki's general solution, neither of the new reduction pairs has empirical implications. Together with the representative sentence (7) of the original reduction pairs (8) and (9), these new reduction pairs are equivalent to (8) and (9) and therefore to T.

In the discussion above, the new postulates (13) and (14) for *F* and un-*F* were obtained by first equivalently reformulating *T* so that un-*F* does not appear in the postulates for *F* and vice versa, and then applying Przełęcki's solution to both polar predicates. In order to produce explicit definitions, the solution suggested in response V can be substituted for Przełęcki's. That is, *T* can be equivalently reformulated by replacing *F* and un-*F* in postulate (3) by their definientia, which leads to sentence (7). Applying response V to this new formulation of *T* leads to the explicit definitions

(15) $\forall x\{Fx \leftrightarrow [\varphi(x) \land \neg\psi(x)]\}$

and

(16) $\forall x\{\text{un-}Fx \leftrightarrow [\psi(x) \land \neg\varphi(x)]\}$.

Together with the empirical claim (7), these new definitions are equivalent to *T*, and since explicit definitions are conservative (Belnap 1993), they do not have empirical implications. They also entail (3) (the postulate that *F* and un-*F* are contraries) and the reduction pairs (13) and (14), so the conjunctions of (3) with the new definitions and (3) with the new reduction pairs do not have empirical implications either.

Note that *T* entails (13) and (14) but not the new definitions (15) and (16). Generally, substituting these reduction pairs for the original definitions in *T* therefore reduces the conceptual commitment, while substituting the new definitions changes it.

3. New Postulates for "Good" and "Bad"

The new reduction pairs (13) and (14) and the new definitions (15) and (16) avoid all of Alfano's criticisms. First and foremost, the conjunctions of these postulates with postulate (3) are consistent with Alfano's core claim (4) and any future empirical results (rebuttal IV). Therefore the postulates do not assume dialetheism (rebuttal I) and do not require abandoning postulate (3) (rebuttal II). These postulates also do not force an arbitrary choice between taking "good" to be prior to "bad" or vice versa (rebuttal V) because the postulates for *F* do not contain un-*F* and vice versa. Furthermore, changes to the original definitions are symmetric in the sense that they are invariant under the simultaneous swapping of *F* and un-*F* as well as φ and ψ. Finally, the postulates do not make the polar predicates into contradictories (rebuttal VI).

One might criticize the new reduction pairs (13) and (14) for failing to fully address rebuttal III. This is because for some objects, it is not determined whether they are *F* or not *F* (un-*F* or not un-*F*) – or in this case, good or not good (bad or not bad). There are two responses. The first is to bite the bullet and accept that "good" and "bad" are (first order) vague:

Some things are clearly good, some things are clearly not good, and for some things, it is not clear whether they are good or not good. This may simply be a fact about the predicates, but it may also be a lack of knowledge. Indeed, Alfano objects to changing the definientia φ and ψ so that claim (4) is false (rebuttal IV) because such a change may lead to empirical objections in the future. In light of this worry, it may be prudential to keep the predicates undetermined for cases in which not enough is known.

For response-dependence theory, this response means that it is a fact of either language or our knowledge that whenever someone is disposed to have a positive sentiment toward x and someone is disposed to have a negative sentiment toward x, it is not clear whether x is good, bad, not good or not bad. For fitting-attitudes theory, this situation occurs whenever it would be fitting to take an approbative attitude, but also fitting to take a disapprobative attitude toward x. In this response, then, φ and ψ become defeasible indicators of goodness and badness, respectively. When the indicators for goodness and badness both apply in the same instance, all bets are off.

The first response is unsatisfactory if fitting-attitude and response-dependence theories are intended to provide complete reductions of "good" and "bad" because the two reduction pairs do not entail that "good" and "bad" are contraries, and therefore the purported reductions fail to reproduce all of the predicates' properties. This problem is solved by the second response, which consists in using the explicit definitions (15) and (16) instead of (13) and (14). This response leads to a very exclusive notion of "good" and "bad." In response-dependence theory, it means that if anyone is disposed to have a negative sentiment toward x, it is not good, and if anyone is disposed to have a positive sentiment toward x, it is not bad. The situation in fitting-attitudes theory is analogous.

The new postulates also avoid the conceptual inconsistency to which the original definitions (1) and (2) give rise if claim (7) is not a conceptual truth. The new definitions (15) and (16) logically entail postulate (3) independently of the status of (7); this makes (3) a conceptual truth, as intended. The reduction pairs (13) and (14) do not entail (3), but are consistent with (3) being conceptually true while (7) is conceptually contingent, because the conjunction of (3), (13) and (14) is conservative. On the other hand, claim (7) entails $\forall x\{[\varphi(x) \wedge \neg\psi(x)] \leftrightarrow \varphi(x)\}$ and $\forall x\{[\psi(x) \wedge \neg\varphi(x)] \leftrightarrow \psi(x)\}$. So if (7) is conceptually true, then for all x, $\varphi(x) \wedge \neg\psi(x)$ is equivalent to $\varphi(x)$ and $\psi(x) \wedge \neg\varphi(x)$ is equivalent to $\psi(x)$ for conceptual reasons. Therefore substituting the new postulates for the original definitions (1) and (2) does not lead to a conceptual change.

Without the assumption of Alfano's empirical claim (4), the argument of this note works as follows: Response-dependence and fitting-attitudes theories have a structure T that entails (7), which either is or is not a conceptual truth. If (7) is *not* a conceptual truth, then the original definitions (1) and (2) cannot both be conceptual truths. Since response-dependence and fitting-attitudes theories claim the conceptual truth of (1) and (2), they are false. The new reduction pairs (13) and (14) and the new definitions (15) and (16) can be conceptually true even if (7) is not, and insofar as they avoid all of Alfano's rebuttals I–VI, they are acceptable substitutes for the original definitions (1) and (2). If (7) *is* a conceptual truth, substituting the new postulates for the original definitions does not amount to a conceptual change. In either case, then, substitution of the new postulates is acceptable. In some cases, it may even save response-dependence and fitting-attitudes theories from contradiction.

4. Conclusion

Alfano concludes that response-dependence and fitting-attitudes theories are false because they have false empirical implications, and that neither theory can be salvaged. In this note I have argued, first, that both theories are false if these implications are conceptually contingent (even if they are in fact true), and second, that the two theories can be saved. Specifically, by adopting new postulates for the polar predicates "good" and "bad" that either entail or are consistent with the postulate that the two predicates are contraries, I have shown that these theories can avoid any non-conceptual implications. Therefore, even if the theories are false in their current formulation, the conclusion should not be that they are untenable, but only that their postulates for "good" and "bad" require modification.

The specific case discussed here points to some general strategies for developing postulates for value notions. If all postulates can be expressed as reduction pairs, empirical and conceptual inconsistencies can be precluded by ensuring that the representative sentences are tautologies. In general, postulates for value notions must be conservative. If value notions are to be completely reduced, then the conceptual truths holding between them must be entailed by the reducing postulates (whether they are explicit definitions, reduction pairs or of some other form).

There is also a meta-philosophical conclusion: The success of the methods I have employed to arrive at new postulates for "good" and "bad" shows that some results from the theory of concept formation are applicable outside of their original domain. This is unsurprising, for the theory of definition has long been considered to be analogous to the theory of inference (see, for example, Belnap 1993), and so methods to improve concepts, just as methods to improve arguments, should be expected to be useful in a wide variety of cases.[5]

Sebastian Lutz
Utrecht University
Theoretical Philosophy Unit
sebastian.lutz@gmx.net

[5] I thank Thomas Müller, Alana Yu, and an anonymous referee for helpful comments.

References

M. Alfano. 2009. “A Danger of Definition: Polar Predicates in Metaethics,” *Journal of Ethics & Social Philosophy* 3(3), www.jesp.org.

N. Belnap. 1993. “On Rigorous Definitions,” *Philosophical Studies*, 72(2/3): 115-146.

R. Carnap. 1936. “Testability and Meaning,” *Philosophy of Science*, 3(4): 420-468.

M. Przełęcki. 1961. “Theoretical Concepts and Experience,” *Studia Logica*, 11(1): 135-138. Summary of M. Przełęcki, 1961. “Pojęcia Teoretyczne a Doświadczenie.” *Studia Logica*, 11(1): 91-129.

W. Rabinowicz. 2008. “Value Relations,” *Theoria*, 74(1): 18-49.