Locating the Wrong Kind of Reason Problem

The *buck-passing account of value* (BPV) attempts to analyze value in terms of reason and pro-attitude: for example, the fact that Marie Curie is admirable can be explained by the fact that some facts about her are reasons to admire her.

BPV faces a well-known difficulty: the *wrong kind of reason problem* (the WKR problem). Imagine a powerful demon threatens to torture you unless you admire him. The fact that admiring the demon will save you from being tortured is, intuitively, a reason to admire him. This fact, however, doesn't make the demon admirable. For BPV to succeed, therefore, it must distinguish the right kind of reason from the wrong kind. Unfortunately, no solution hasn't been widely accepted yet.

The WKR problem isn't new, however. BPV is a version of the *fitting-attitude analysis of value* (FA), which attempts to analyze evaluative terms or facts in terms of deontic terms or facts: for example, the term "being admirable" is analyzed in terms of "being correct (fitting, appropriate) to admire". However, FA faces a problem similar to the WKR problem (D'Arms & Jacobson, 2000a, b; Ewing, 1959). Take the demon scenario for example. It would be *prudentially correct* for you to admire the demon, but it is *incorrect*—for the purpose of analyzing "being admirable"—to admire the demon. Accordingly, FA must distinguish between the right kind and the wrong kind of correctness. Call it the *wrong kind of the deontic problem*.

Given the wrong kind of the deontic problem, it's tempting to think that we should give up FA. Nevertheless, I argue that this view is wrong. Let's consider how to analyze, say, "admirable" in terms of other evaluative terms since it's unlikely that it is unanalyzable. Consider a plausible account of "admirable" in terms of evaluative terms:

(WORTH) X is admirable, if and only if X is *worthy of admiration*.

WORTH has many virtues. First, it's applicable to other evaluative terms: e.g., "desirable" can be understood as "worthy of desire" and "enviable" as "worthy of envy". Second, WORTH is analytically true.

However, a problem similar to the WKR problem also afflicts WORTH. Again, in the demon scenario, it seems correct to say that "the demon is worthy of your admiration because doing so will save you from being tortured". We may call it "the *wrong kind of worthiness problem*". Replacing "being worthy of" with other evaluative terms, such as "deserving" or "meriting", doesn't help. For it seems correct to say that "the demon *deserves* or *merits* your admiration because doing so will save you from being tortured".

The wrong kind of worthiness problem shows that the WKR problem lies in *analysandum* rather than analysans. So, FA should not be rejected because of the WKR problem. Since it's unlikely that evaluative terms, such as "admirable", "enviable", "funny", are unanalyzable, the wrong kind of worthiness problem is every

value theorist's problem.

How to solve the wrong kind of worthiness problem? The solution, I think, can be found if we look more closely into the claim, "the demon is worthy of your admiration because doing so will save you from being tortured".

First, we can notice that the demon is worthy *of your admiration, but not others' admiration*. One may thus think that Schroeder (2010) is correct that the right kind of reason is the one necessarily shared by everyone who engages in the activity of admiration, and the wrong kind of reason is the one valid only for some people. However, Schroeder's solution has a flaw. While the fact that the demon is worthy only of your admiration rules itself out as the right kind of reason to admire the demon, it may suggest that the demon is *admirable to you*. Schroeder's solution needs to say more about how to deal with this kind of agent-relative value.[1]

Second, we can notice that the demon is worthy of your admiration *because doing so will save you from being tortured*. It shows that the wrong kind of reason to admire the demon is the one that appeals to the *benefit* or *consequence* of admiring the demon (Lang, 2008; Rowland, 2013). So, the right kind of reason is the one regardless of the benefit or consequence of admiring the demon.

Accordingly, we may distinguish two kinds of worthiness: *worthiness for one's own sake* and *worthiness for the others' sakes*. WORTH can be modified as follows:

> (RK-WORTH) X is admirable, if and only if X is worthy of admiration *for its own sake*.

When X is worthy of admiration for its own sake, it is worthy of admiration for its own feature regardless of any benefit or consequence of being admired. And if X is worthy of admiration for the others' sakes, e.g., for the benefits or consequences of being admired, X is not admirable on such grounds.

Therefore, the WKR problem is ultimately the problem of wrong kind of worthiness. It is not simply a problem for FA. Moreover, my discussion supports Lang and Rowland's solutions to the WKR problem, but it explains why their solutions are on the right track: the right kind and wrong kind of reason pick out two different kinds of value meriting our responses. But the distinction in value is something we are already familiar with: being valuable (worthy, admirable, desirable) for one's own sake and being valuable for the sakes of others. Some objects are worth our certain responses for their own sake, but some others only for the sakes of other

---

[1] Schroeder does discuss the problem caused by some agent-relative value, like "good-for": for example, what is the reason shared by everyone with regard to "good for a wick person"? Schroeder answers that it can be analyzed in terms of the reason shared by everyone who engages in *watching out for the wick person*. It is unclear to me, however, how this strategy can apply here.