I. Origins of the Wrong-Kind-of-Reason Problem, Pt. 1: D’Arms & Jacobson

D’Arms & Jacobson argue that the following “neo-sentimentalist” thesis is accepted by a wide range of philosophers who are otherwise divided in their metaethical commitments (“Sentiment and Value,” p. 729):

**response dependency thesis (RDT):** To think that X has some evaluative property \( \Phi \) (such as being admirable, or being disgusting, or being fearsome) is to think it appropriate (or merited, or fitting) to feel sentiment \( F_\Phi \) (such as admiration, or disgust, or fear) in response to X.

Four comments about RDT:

- Brandt’s influence here should be clear. (D&J were grad students at Michigan, where Brandt taught.)
- D&J add the ‘to think’ on either side of the analysis to make it compatible with expressivism, which aims to provide an account of normative thought (and talk), not of normative properties themselves.
- I think ‘response dependency’ is a highly misleading label for this thesis and is best avoided. RDT does not posit that properties such as being admirable are dependent on our responses; rather, it posits that they are dependent on the fittingness of certain responses we might have. (Are punishable offenses “punishment dependent” because they merit punishment?)
- I disagree with D&J that this thesis is distinctively sentimentalist, for two reasons. First, a similar analysis holds for various normative properties that have nothing to do with the sentiments (such as being punishable, hirable, credible, useful, persuasive, or choiceworthy).

Second, RDT is compatible with facts about the fittingness of sentiments being themselves grounded in facts having nothing to do with the sentiments.

D&J think many neo-sentimentalists make a mistake, which they characterize in two slightly different ways:

In one article, it is called the moralistic fallacy: the fallacy of inferring <Sentiment F is not a fitting response to its object X> from <F is wrong (or inexpedient) to feel>, or of explaining the truth of the former by appealing to the truth of the latter.

In another article, it is called the conflation problem: the problem of not having the “resources to differentiate . . . moral and prudential reasons for feeling a sentiment . . . from reasons bearing on whether [the sentiment fits its object]” (“Sentiment and Value,” p. 732).

D&J deploy a number of wonderful examples to demonstrate how this problem arises:

- The fact that envying your newly tenured colleague, Susan, might jeopardize your relationship with her and hence undermine your chances of getting tenure yourself is a reason not to envy her, but it does not bear on whether she is enviable (“Moralistic Fallacy,” p. 71).
- The fact that the wolf is more likely to attack you if you are afraid is a reason not to fear the wolf, but it does not bear on whether the wolf is fearsome (ibid., p. 87).
- The fact that a joke is cruel or offensive to a socially marginalized person or group is a reason not to be amused by it, but it does not bear on whether the joke is funny (“Sentiment and Value,” p. 731).
- The fact that being ashamed-of-one’s-inability-to-articulate-one’s-philosophical-views-clearly will inhibit one from speaking up in seminar, thus exacerbating the problem, is a reason not to be ashamed of being an inarticulate philosopher, but it does not bear on whether being so is shameful (ibid., p. 741).
The basic problem: not every reason for/against feeling sentiment $F_\Phi$ toward object $X$ is a reason in virtue of which $X$ is or is not $\Phi$. So defenders of RDT need a way of circumscribing the reasons for/against feeling $F_\Phi$ toward $X$ that do bear on whether it is fitting to feel $F_\Phi$ toward $X$ from “the wrong kind of reason” (“Sentiment and Value,” pp. 734, 735) that do not.

When D&J talk about reasons that “are (ir)relevant to” or “that do bear on” whether $X$ is $\Phi$ (or whether it is fitting to feel $F_\Phi$ in response to $X$), they seem to be talking about whether or not these reasons ground the fact that $X$ is $\Phi$ (or the fact that it is fitting to feel $F_\Phi$ in response to $X$).

So they seem to be assuming that the fact that it is fitting to feel $F_\Phi$ toward $X$ obtains in virtue of the balance of reasons (of the right kind) to feel $F$ toward $X$, and hence assuming that every (right kind) reason to feel $F$ toward $X$ is part of what makes it the case that it is fitting to feel $F_\Phi$ toward $X$.

They might also be assuming that any partial ground of the fact that it is fitting to feel $F_\Phi$ toward $X$ is a (right-kind) reason to feel $F$ toward $X$, but that would be a bad assumption to make.

Even though most of D&J’s examples involve moral or prudential wrong-kind reasons to feel $F_\Phi$, the problem can’t be solved by restricting the right-kind reasons to non-moral, non-prudential reasons.

Some right-kind reasons are prudential: the fact that the mouse can’t hurt you is a reason against being afraid of the mouse that bears on whether the mouse is fearsome (“Moralistic Fallacy,” p. 87).

And some right-kind reasons are moral: the fact that Susan deserved to get tenure is a reason against feeling outraged at her getting tenure that bears on whether her getting tenure is outrageous (ibid., p. 82).

D&J’s own solution to the conflation problem, as I understand it:

Each emotion has a given shape: it “presents its object as having certain evaluative features” (p. 73).

For example, envy presents (i) its object as being a rival, (ii) this rival as having something good you don’t have, and (iii) it being bad for you that the rival has this thing and you don’t.

Each emotion also has a size: its intensity or strength. (Felt strength? Propensity to result in action?)

For example, outrage has a larger size than anger, which has a larger size than annoyance.

An emotion is fitting, insist D&J, if and only if, and because, (a) things are as its shape presents them to be, and (b) its size is not an overreaction.

So right-kind reasons either bear on whether things are as F presents them to be or bear on whether F is an overreaction.

Three reservations about their solution:

1. Since they appeal to evaluative categories such as goodness and badness in the content of an emotion’s presentations, either D&J cannot apply RDT to those evaluative categories (and hence are not at bottom neo-sentimentalists), or they must find some sentiments whose presentations have entirely non-normative contents, or they must be fine with explanatory circles.

2. We cannot apply their solution to a normative property, $\Phi$, whose corresponding $F_\Phi$ does not have a shape. So how do we avoid the conflation problem for being punishable, hirable, credible, persuasive, etc.?

3. ‘Overreaction’ is a normative term, and presumably a sentiment is an overreaction in virtue of the right-kind reasons for and against it. So the conflation problem appears to rearise when we assess whether a sentiment is unfitting due to its size.

One way out of this last problem would be to stop distinguishing size from shape, and to fold a sentiment’s size into its shape. For example, maybe envy to degree $d_1$ presents a rival as having something good to degree $d_2$ that you don’t have, where it’s bad for you to degree $d_3$ that this is so.
II. Origins of the Wrong-Kind-of-Reason Problem, Pt. 2: Rabinowicz & Rønnow-Rasmussen

According to Rabinowicz & Rønnow-Rasmussen, versions of the following thesis have been defended by Brentano, Broad, Ewing, McDowell, Chisholm, Falk, Gibbard, Anderson, Lemos, Scanlon, and many others:

the fitting-attitudes analysis of value: For it to be the case that X is valuable (or good, or Y-able) is for it to be the case that it is fitting for any suitable agent (or any suitable agent ought, or there is sufficient reason for any suitable agent, or there is pro tanto reason for any suitable agent) to have certain pro-attitudes or positive emotions toward X.

Different theorists vary with regard to (i) the type of normative property for which they apply this analysis, (ii) whether the relevant normative status concerns fittingness, oughtness, or reasons, (iii) how they determine the range of suitable agents, and (iv) how they determine the relevant pro-attitudes or emotions.

Here is how R&R hold we can make sense of the final vs. instrumental value distinction, on this analysis:

For it to be the case that X has final value is for it to be the case that it is fitting for any suitable agent to have certain pro-attitudes toward X for its own sake.

For it to be the case that X has instrumental value is for it to be the case that it is fitting for any suitable agent to have certain pro-attitudes toward X for the sake of its effects.

R&R interpret Scanlon’s buck-passing account of value as being as version of the fitting-attitudes analysis, like so:

the buck-passing account of value (BPV): For it to be the case that X is good is for it to be the case that there is sufficient reason for any suitable agent to have certain pro-attitudes toward X.

They then mostly focus on this as their canonical version of a fitting-attitudes analysis of value.

Three comments:

- This is not the best interpretation of Scanlon. Among other problems, Scanlon appeals to reasons for certain actions (such as protecting X) as well as reasons for certain pro-attitudes in his account.

- Gertken & Kiesewetter use ‘valuing’ as an umbrella term covering the relevant pro-attitudes and positive emotions, but this is problematic, because valuing is itself a pro-attitude, and it is not clear that all the relevant pro-attitudes and emotions are determinates of the determinable valuing.

- It seems to me that it is positively misleading to call versions of this analysis that appeal to fittingness, oughtness, and reasons all by the same name, since these are all very different normative categories. ‘Fitting’ is a bad umbrella term to use here, especially if fittingness is not a deontic notion.

R&R raise a challenge for the BPV that they call ‘the wrong-kind-of-reasons problem’, but which is basically the same as D&J’s conflation problem. The problem is that the BPV, as formulated, appears to be extensionally inadequate, because of cases like the following:

1. An evil demon threatens to inflict severe pain on us unless we desire a certain saucer of mud; thus we have sufficient reason to desire the saucer, but this doesn’t mean the saucer of mud has value.

2. According to most hedonists, knowledge, freedom, and friendship don’t have final value, but we all have sufficient reason to favor (to desire, strive for, etc.) these things for their own sake.

3. According to some mental-state axiologies, only mental states have final value, and we have sufficient reason to admire or cherish certain non-mental objects for their own sake.

Thus we need to distinguish two types of reasons: reasons “of the right kind” to plug into the BPV and reasons “of the wrong kind” to plug into that account. In effect, we have revised BPV as follows:

BPV*: For it to be the case that X is good is for it to be the case that, setting aside reasons of the wrong kind, there is sufficient reason for any suitable agent to have certain pro-attitudes toward X.
the wrong-kind-of-reason problem (for the BPV): How do we provide an account of the distinction between the right and wrong kind of reasons that makes BPV* extensionally adequate and non-circular?

Three solutions to this problem that R&R argue fail:

1. **state-given vs. object-given reasons:** Maybe wrong-kind reasons are “state-given reasons” for a pro-attitude in that they are “provided by properties of the attitude itself,” whereas right-kind reasons are “object-given reasons,” in that they are “provided by properties of the [pro-attitude’s] object” (pp. 404–5).

**R&R’s initial objection:** Properties are cheap: if pro-attitude \( x \) toward object \( a \) would have property \( P \), then, ipso facto, \( a \) has the property \( P’ \) of being such that attitude \( x \) toward it would it would have \( P \). “Consequently, to the attitude-given reason, provided by \( P \), corresponds the object-given reason, which is provided by \( P’ \ldots \)” (p. 406).

But this last inference is fallacious. Just because \([x \text{-toward-} a \text{ would have } P]\) provides a reason, it doesn’t immediately follow that \([a \text{ is such that } x \text{-toward-it would have } P]\) provides a reason.

**R&R’s main objection:** The proposal doesn’t work against a variant of the demon case in which the demon threatens to inflict severe pain on us unless we admire him (rather than the saucer of mud).

2. **wrong-kind-of-reason skepticism (WKR skepticism):** Maybe all putative wrong-kind reasons to have pro-attitude \( A \) are really reasons to want (or to try, or to intend) to have \( A \), or reasons to bring it about that one has \( A \). Since all the reasons for \( A \) itself are of the right kind, the wrong-kind-of-reason problem does not arise.

**R&R’s first objection:** Following D&J, they hold that WKR skepticism leads to “deeply unacceptable” results. For example, in our demon case, “were [my] will effective” in shaping my attitudes, I would fail to admire the demon while desiring in vain that I admire him (“Sentiment and Value,” p. 744).

**R&R’s second objection:** WKR skeptics like Gibbard and Parfit have no argument for their view.

But this is false. According to one popular argument due to Kelly, Parfit, Shah, and others, wrong-kind reasons for pro-attitude \( A \) don’t exist because we can’t base \( A \) on those reasons.

**R&R’s third objection:** WKR skepticism only shifts the problem, since WKR skeptics owe us an account of “what makes something a reason for wanting (or trying) to have a certain attitude toward an object rather than a direct reason for having the attitude in question” (“Strike of the Demon,” p. 414).

**Gertken & Kiesewetter’s reply** (“Right and Wrong Kind of Reason,” p. 11, n. 25): Advocates of the BPV who are not WKR skeptics also need to explain this distinction (if it even needs explaining), so this is a problem that is independent of the wrong-kind-of-reason problem (if it is even a problem).

3. **R&R’s dual-role proposal:** Maybe the fact that object \( X \) has property \( P \) is a right-kind reason with regard to pro-attitude \( A \) when (and only when) it counts in favor of possessing the following complex attitude: \((A \text{-toward-} X)\text{-because-}(X \text{-has-} P)\).

(Note that now we need to reformulate BPV* slightly, since on this proposal right-kind reasons do not favor pro-attitudes as such, but rather favor pro-attitudes-on-certain-bases.)

**Tersman’s objection:** Suppose the demon threatens to inflict severe pain in us unless we admire him “precisely on . . . account of his determination to punish us if we don’t” (“Strike of the Demon,” p. 419).

R&R conclude that the wrong-kind-of-reason problem is an unsolved challenge to the buck-passing account, and to fitting-attitude analyses of value more generally.

**III. Subsequent Work on the Wrong-Kind-of-Reason Problem(s)**

D&J’s and R&R’s articles have spawned a massive literature on the wrong-kind-of-reason problem.
G&K provide a helpful overview of other solutions that have been put forward (p. 3):

- **derivative vs. non-derivative reasons:** Maybe wrong-kind reasons are instrumental or derivative reasons for attitudes (that is, “reasons for responses that one has because that response is conducive to something else one has reason to do or want”), whereas right-kind reasons are not.

- **consequence-dependent vs. consequence-independent reasons:** Maybe wrong-kind reasons “depend on the benefits or, more generally, consequences of the attitude,” whereas right-kind reasons “are provided by features that would also provide reasons in the absence of these benefits, or independently of any consequences.”

- **correctness-independent vs. correctness-related reasons:** Maybe wrong-kind reasons “are independent of the correctness of the relevant pro-attitude,” where right-kind reasons “are essentially related to the correctness of the pro-attitudes, e.g. by suggesting that the attitude meets its internal standard of correctness.”

- **idiomsyncratic vs. shared reasons:** Maybe right-kind reasons for some activity are necessarily shared by everyone engaged in that activity because they are engaged in that activity, whereas wrong-kind reasons are not.

Needless to say, none of these proposals has gained wide acceptance.

Another important issue that has arisen in the literature on this topic is the degree to which the right- vs. wrong-kind of reason distinction can be drawn for reason-directed entities other than pro-attitudes and emotions, such as:

- **beliefs:** The distinction in the ethics-of-belief literature between *epistemic reasons for belief* (i.e. reasons for belief provided by one’s evidence, or by the truth-conduciveness of process by which the belief would be formed, or by whatever one’s preferred ground of epistemic notions is) and *practical reasons for belief* (e.g. reasons provided by the benefits to oneself or others of holding the belief) feels very close to the distinction between right-kind and wrong-kind reasons for pro-attitudes.

- **intentions:** In Kavka’s toxin puzzle, the reason you have to intend drink the toxin provided by the billionaire’s offer seems similar to a wrong kind of reason for a pro-attitude, and the reason you have against intending to drink the toxin provided by the discomfort it will cause you seems similar to a right kind of reason for a pro-attitude.

- **actions:** Some authors hold that the right- vs. wrong-kind distinction does not apply to reasons for action, whereas other authors hold that it does. For instance, Schroeder has argued that there is a right- vs. wrong-kind distinction for any activity governed by internal standards, including tying a knot, executing the Lucena endgame in chess, and setting the table for a White House state dinner.

Those who hold that the right- versus wrong-kind reason distinction extends to some or all of the above cases often claim that the wrong-kind-of-reason problem is a problem for more than just advocates of buck-passing, fitting-attitudes, or neo-sentimentalist analyses of value.

- **first alleged upshot:** “The wrong-kind-of-reason problem is everyone’s problem.”

- **second alleged upshot:** Alleged solutions of the problem that cannot be generalized to the case of belief, intention, and/or action are not fully adequate responses to the underlying problem.

But, as G&K note, it is not entirely clear what we mean by ‘reason of the wrong kind’ once we attempt to generalize the problem in this way, nor is it clear that the supposed more general problem is the just same problem with a larger focus.