8

Extrinsic Value and the Separability of Reasons

BARRY MAGUIRE

SECTION ONE: THE PUZZLE

It is the Final of the World Cup in 1966. Pelé is running with the ball around England players, who by comparison seem hardly to move. He started running near his own penalty box, and he has already passed three players, four, five. Two minutes left and the game is tied. There are one hundred and twenty thousand Brasilians in the stadium, and another one hundred and fifty million rambunctiously shouting at televisions.

You are on the roof of the stadium, undetected, with a sniper rifle. Seven players now, eight. Pelé is bearing down on the last couple of defenders. You check wind speed and direction for the final time, cock the trigger, your crosshairs are following the middle of Pelé's back. That's him in front of the keeper now, and you have a full half-second in which to gently squeeze the trigger.

* * *

Here's the problem. You are an Act Consequentialist. You think that one ought to perform whichever option in any given situation would maximize final (i.e. non-instrumental) value.

The premature death of a footballing hero on an international stage would be a terrible thing. However, there are millions of people each of whom would be deeply saddened if Pelé were to die suddenly on the field. It is natural to think that there is something valuable about such grief in the face of such a tragedy. For imagine if people were utterly indifferent, or worse, pleased by such an event. Since there are millions of such reactions, the death of Pelé would bring about millions of valuable states of affairs. Now, surely the absolute value of any one instance of grief at Pelé's death will be less than the disvalue of the death. So the pressing question for the Act Consequentialist (henceforth:

Consequentialist) is this: will the overall value of the consequences—the disvalue of Pelé's death together with the value of all these virtuous reactions—be positively valuable?

So long as these values aggregate in some non-exotic way, this will just be a question about the numbers of people involved. So let's add that you accept the following plausible principle concerning the aggregation of values, that each additional negative reaction to Pelé's death adds some separate, non-diminishing amount of value to the overall value of the consequences of shooting Pelé. Then, so long as there are enough people watching the game, the value of all these reactions will be absolutely greater than the disvalue of shooting Pelé.

There are some other values at stake. Plausibly in addition to the value of a negative response to a disvaluable state of affairs there will be various associated disvalues: the disvalue of the loss of Pelé's company on future family occasions, the disvalue of lower quality football on television, and so on. Again, it seems plausible that enough valuable reactions will eventually swamp any of these disvalues.

We assume that you, the would-be sniper, have nothing better to do this afternoon. All of your other options would promote zero net value or less, so this option would maximize value as long as the overall value is greater than zero, as indeed it would be. These claims together with your Consequentialism entail that you ought to shoot Pelé. You are a strong willed Consequentialist. You shoot him.

* * *

This isn't the first time your morals have led you astray. You have acted in the past on similar reasoning. If you kick a dog in a crowded street many people will suffer empathetically. Their suffering empathetically is valuable. So, with an iron will, you go around kicking dogs on busy streets. For a while you carried around a razorblade, with which to slice open the bottom of people's shopping bags in crowded areas, thereby to allow passersby to manifest virtue by helping to pick up runaway potatoes. Soon you realized that you could simply trip elderly people to the same effect. You have even, from time to time, deliberately injured yourself, thereby to manifest courage, resilience, and fortitude by dealing with the pain and quickly adjusting to the inconvenience.

But clearly you should not have done these things. It is not the case that you ought to shoot Pelé in this situation, or trip up elderly people in busy streets. These are substantive ethical claims. But they seem utterly

uncontentious. So we have a contradiction. At least one of these claims or assumptions is false.

SECTION TWO: SOME PREMISES DEFENDED

We shall accept the non-normative features of the set-up. You have the rifle; you are on the roof; there are millions of people who would grieve. Let's also accept the various non-theoretical axiological and ethical assumptions: the assumption that the death is disvaluable,[1] and that you ought not to shoot Pelé. These two assumptions are independently plausible. The first will be entailed by almost any axiology. The second has overwhelming support from folk morality, or at least from football fans. In this section I will defend the remaining assumptions: that the sadness would be valuable (2.1); that the value of sadness at an event is less than the value of the event (2.2); that the value of the individual sad responses aggregate non-exotically (2.3); and that other things are indeed equal in the Pelé case (2.4).

2.1 The Value of Sadness

Let's start with the claim that grief in response to genuine tragedy is valuable. A little more carefully, consider the claim that grief in response to genuine tragedy for its own sake is valuable. The "for its own sake" refinement—here picking out a psychological feature not a distinction in value—rules out the value of "instrumental" grief: e.g. grief in response to the death on the grounds that one will not thereby be able to kill Pelé oneself tomorrow. This more carefully stated claim has a great deal of independent plausibility.

Start with the thought that sadness at tragedy for its own sake is better than delight at tragedy for its own sake (I'll henceforth assume and suppress the "for its own sake" qualification). That latter attitude would be grotesque. It is also plausible that sadness at tragedy is better than indifference.

Let these claims be granted. They do not quite entail that sadness in response to tragedy is valuable. For it is possible that sadness in response to a disvaluable event is better than happiness or indifference in response to a disvaluable event, but that all three are disvaluable. Perhaps in

[1] This point will go through without taking a stand on whether the premature death of someone with a good life is intrinsically or extrinsically finally valuable.

response to a disvaluable event, there are no positively valuable reactions available? Here are two arguments against the plausibility of this.

First consider regular virtuous responses to unfortunate situations, for instance when a group of strangers console someone who has stumbled on some subway steps. These expressive actions seem to have value in themselves. Indeed, I think this is a really important sort of value—it is one of the beautiful and often overlooked features of life in a large bustling city like London or New York. However, if virtuous expressive actions in response to unfortunate events have value, then surely virtuous attitudes (in situations where there is nothing that can be done) also have value.

Secondly consider the converse claim. Could it be that a negative reaction to a valuable event is worse than a positive reaction but still valuable in itself? Could it be that in response to a valuable event, there are no disvaluable attitudes available? Surely not. We disdain Scrooge for his grumbling about Christmas parties. A negative reaction to a valuable event is disvaluable. Suppose we can assume the relevant kind of symmetry between relevantly valenced attitudes to valuable and disvaluable events. Then we can conclude that a negative reaction to a disvaluable event is valuable.

These axiological claims admit of further support in the form of Thomas Hurka's value-based theory of virtue (2000). Thomas Hurka argues that positive reactions to valuable states of affairs are valuable, negative reactions to disvaluable states of affairs are valuable, positive reactions to disvaluable states of affairs are disvaluable, and negative reactions to valuable states of affairs are disvaluable.[2] For instance, there is value in feeling sad when someone you love dies. There is value in feeling pleased when you succeed in some difficult venture. It is disvaluable to feel unhappy when others thrive, or happy when they are suffering, and so forth.[3] The plausibility of Hurka's theses provides

[2] Bear in mind we don't need to accept the strong thesis that only the value of the object determines (modulo features of the attitude such as strength and duration) the value of the attitude. Other factors may also be relevant to the value of the attitude, for instance facts about the agent's history with the object (i.e., perhaps the object is the person's husband). We merely insist that there is a function of the one to the other.

[3] In each case some particular relevantly valenced attitude, to some degree, is "fitting." This appeal to fittingness presents no difficulties for our provision of a "value-first" normative theory, since fittingness is not *genuinely* normative. Attitudes such as despair, envy, or a murderous rage might be fitting in a situation, though nothing speaks normatively in their favor; one would not be in any way criticizable for failing to have an attitude just in virtue of its being fitting.

support in turn for the claim about the value of sadness at tragedy. It is important to notice, however, that the plausibility of the claim does not depend on the plausibility of Hurka's stronger claims.

Hurka's principles also help to explain where a more familiar thesis goes wrong. According to NAÏVE HEDONISM, only positively valenced attitudes are valuable and only negatively valenced attitudes are disvaluable. We are all familiar with counterexamples to this thesis: sadistic or masochistic silly pleasures, the pain constitutive of athletic achievement, and so on. Hurka's principles are just as systematic as the more familiar NAÏVE HEDONISM, but even more powerful. Valuable or disvaluable attitudes—yours or mine—can themselves be the object of positive or negative reactive attitudes; the principles we have are sufficient to explain the virtue of such higher-order attitudes.

Another kind of hedonism is compatible with Hurka's principles. Let RESTRICTED HEDONISM be the thesis that non-intentional painful experiences are disvaluable and non-intentional pleasant experiences are valuable. Allow that we can distinguish a non-intentional painful experience within the psychology of the grieving football fan. Then even if we allow that the grief at the death is valuable, we must also admit that the painful aspect of the grief is disvaluable. It is then an open question whether the overall final value of any one instance of suffering is positive or negative. This line of thought causes trouble for the assumption in the set-up of our puzzle that there were no other significant values. If every valuable reaction is accompanied by an equally disvaluable experience, our puzzle disappears.

There are various replies available. Perhaps our intuitions here are misled by some inertial pull towards NAÏVE HEDONISM. Perhaps the pain is bad *for* the griever, but not disvaluable. We can deny that all increases in welfare are valuable and all decreases in welfare are disvaluable. Reflecting on the axiological status of sadistic pleasure or retributive punishment can help to motivate this denial. Perhaps the disvalue of a painful experience is a function of the value of the intentional state of which it is a part.

However, I'm inclined to put most weight on our original intuition. When we say that happiness about a tragedy for its own sake would be worse than sadness at tragedy for its own sake, we don't merely mean that it would be worse in a respect, though perhaps better overall. We really mean it would be worse overall. It is a further question whether this value is mitigated by the non-intentional value of the pleasure or pain. But the only thing needed for current purposes is the fact that the

net value of a negative attitude towards disvalue would be positive and that the net value of a positive attitude towards disvalue would be negative. Then so long as we have enough reactions, we still have our puzzle.

2.2 Moore's Thesis

Consider briefly the thesis that the value of an appropriate reaction to an event is absolutely less than the value of that event. Both Tom Hurka (2000) and G. E. Moore (1903) defend a fully general version of this thesis. Moore doesn't offer much of an argument for this thesis, taking it to be fairly evident on reflection. Hurka discusses many examples in its support. He notes that it isn't plausibly better to feel malice together with shame about one's malice than not to feel either; nor worse for one to feel pleasure and another to enviously wish that the one wasn't feeling that pleasure than for there to be no pleasure at all.

For current purposes we don't need to accept the fully general version of this thesis. There doesn't seem to be anything wrong with a training exercise that leaves well-prepared soldiers in the middle of a mountain range to find their own way home. Similarly, perhaps some virtuous episodes are so fabulously intense or long lasting as to outdo whatever they are responses to. But we should accept the thesis in Pelé's case, and ones like it. Since my primary interests here concern the structure of the puzzle itself and the opportunities it affords to showcase underexplored resources of value-first theories, I shall leave the status of the general thesis to the reader to consider.

2.3 The Aggregation Principle

Consider now the thesis that each additional negative reaction to Pelé's death adds some separate, non-diminishing amount of value to the overall value of the consequences of shooting Pelé. Call this thesis VALUE SEPARABILITY. There is the following simple argument for this thesis. The overall state of affairs consisting in the death and all the sadnesses will be non-positive only if the absolute value of the sadnesses sums to less than the disvalue of the death. Therefore, in order to reject VALUE SEPARABILITY one needs to defend a limit on the amount of value that can be summed altogether by additional reactions. But there is no such limit on the *number* of possible reactions, nor any need for reactions to be causally connected to each other in any way. And it is implausible

that the value of one virtuous response to something is less in virtue of some further causally unconnected others having a similar response. Furthermore, the value of each reaction is non-negligible. What's more we are assuming quite generally that values, and hence the value of reactions, are commensurable.[4,5]

VALUE SEPARABILITY is also quite plausible. Perhaps the world is better overall with some of these actions and responses than with none of them. This is clearest in ordinary situations in big cities when a group of people rushes to the aid of someone who has fallen, etc. Much value is promoted in virtuous responses to unfortunate situations and it seems quite possible in principle for the net value of the outcome of such a situation to be positive. This must be how the arguments of the *Theodicies* got their initial plausibility. It is plausible that in *some* cases some "evil" is a necessary condition of certain goods such that the overall outcome is evaluatively positive. It just isn't true of *all* evils.

2.4 On Whether Other Things are Equal

Finally to the assumption that you have no better alternatives to shooting Pelé. An interesting issue here is whether it is possible for you to bring about the virtuous suffering at a lower cost. Could you make it look to the world as though Pelé had been shot even though he was in fact removed through a trapdoor? We assume that negative reactions based on justified but false beliefs would still have value.[6] This would be evaluatively preferable to actually shooting Pelé. Unfortunately

[4] I acknowledge a hostage to fortune here. It is possible that an axiologist might develop a "lexical priority" view, perhaps along the lines of J. S. Mill's higher/lower pleasures distinction, or Rawls' distinction between the deontic significance of his first and second principle of justice. Such a view would help with the Pelé case, as well as T. M. Scanlon's "electrocution" case (1998: 235), along with an excellent challenge offered to me by Doug Portmore: wouldn't it obviously be a disvaluable thing overall if Pope Francis died tomorrow, no matter how many people would grieve? The lexical priority view depends on a distinction between kinds of values (e.g., death, pleasure, virtue) rather than kinds of value (intrinsic, extrinsic). In addition to the argument in the main text, the latter seems better suited to the structural distinction—non-separability—that we are discussing in this paper. I would welcome developments of the more familiar lexical account in this context. For now I stick with Hurka.

[5] I am also assuming a plausible non-satiation thesis about value: adding value always increases the overall amount of value.

[6] Hurka devotes a chapter to complications arising from false beliefs about events that could have happened, and true beliefs about fictions. In short his view is that the value of the attitude doesn't vary with the truth of the belief when the object is not taken to be fictional, and the value is less when the object is fictional.

(we stipulate) the ground is too hard to build the needed trapdoor, and you couldn't find a decent smoke machine in time.

So. Once we have enough responses, the overall value of the situation may be positive. The trouble arises, or so it seems to me, when we consider *bringing about* one of these sets of states of affairs. Clearly, in many such cases, you shouldn't bring them about, even *if* they are valuable overall—even God should not. In other words—and this is the key point—these kinds of cases motivate us to separate out facts about overall value from deontic facts about what you should do. Reflection on these cases motivates us to look for an alternative to Consequentialism.

SECTION THREE: THE PUZZLE FOR THE VALUE-BASED REASONS FOR ACTION

The structural complexity in the metaphysics of ethics is not respected by the simple exhortation to maximize value in every choice situation. A value-first ethical theorist can move towards a more faithful account of the metaphysics of ethics in two steps.

The first step is to appeal to facts about various relations between actions available to an agent and valuable states of affairs to explain facts about that agent's normative *reasons*. We will find our solution by attending carefully to facts about such relations. As value-first theorists we are free to appeal to various different kinds of structural complexity within axiology, or complexity in the principles relating axiological facts and other ethical facts, or both, as well as to a variety of different relations between actions or omissions and valuable states of affairs, to account for complexities we find within the metaphysics of ethics. It is the purport of this paper to put some of this available machinery to good use.

In the second step—which after this paragraph I shall say no more about—these facts about normative reasons explain strict deontic facts: facts about ought, permissibility, blameworthiness, etc. Facts about individual reasons (count noun) for and against *x*-ing explain how much reason (mass noun) there is to *x*. Hence we can talk about whether there is reason to *x*, meaning whether the net weight of reasons is positive, whether there is weighty reason to *x*, meaning whether the net weight of reasons is positive and weighty relative to some (perhaps contextually shifty) standard, and whether there is most reason to *x*, meaning whether the net weight of reasons to *x* is greater than the net weight of reasons to do

anything other than *x* in the situation in question. Then the simplest thesis relating reasons and oughts maintains that you ought to do whatever you have most reason to do. There are more sophisticated accounts of this relation that provide a distinctive normative role for sufficiency, supererogation, and the like; we don't need to get into those details here.[7] We can assume the simple thesis for our purposes.

Begin with a simplifying assumption, namely that reasons are facts of the form *x*-ing would promote S, where *x*-ing is a particular action performed by a particular agent and S is any state of affairs. Reasons regimented differently are explained in terms of their relations to reasons of this kind.[8] For now: an action *would promote* a state of affairs by (partially or fully) probabilifying it.

Let R be such a fact of the form [*x*-ing would promote S]. The simplest value-based theory of reasons[9] maintains that:

VALUE-REASONS: The fact R that some action *x* available to you would promote some state of affairs S is a reason for you to perform that action if and only if and because S is valuable.

This principle captures a core Consequentialist motivation. I am interested in exploring solutions to the puzzle that begin by accepting this principle. The key feature of *value* is that it is a monadic graded property of states of affairs.[10] For now I want VALUE-REASONS to be as neutral as possible on substantive questions about values and reasons. I allow "value" to range over both instrumental and final value. Final value is the value something has for its own sake. In turn "final value" ranges over *intrinsic* final value, the value something has for its own sake just in virtue of its intrinsic properties, and *extrinsic* final value, the value something has for its own sake partly in virtue of at least one extrinsic property. There are a variety of kinds of extrinsic final value, including part value, symbolic value, conditional value, etc.[11]

[7] For further discussion of this and related issues, see my co-authored (2016).

[8] E.g. as evidence of some such fact or part of the explanation of the obtaining of some such fact. I address other kinds of reasons briefly in footnote 26 and at length in my (2016).

[9] For further discussion, see my (2016).

[10] I assume for simplicity that one state of affairs is better than another if and only if and because the one has more value than the other, and that states of affairs rather than objects are the value-bearers. For discussion see Zimmerman 2001 and Wedgwood 2009.

[11] For further detail, see Ralf Bader's "Two Value-Theoretic Trichotomies" (ms).

We assume that value is a cardinal ordering on states of affairs.[12] We allow for the theoretical possibility that some states of affairs have zero value and some reasons have zero weight. Given this assumption, and since every action would promote a large number of states of affairs, it turns out on this way of setting things up that there are a large number of reasons for and against every action. This is not a *reductio*. In general we are interested not in the existence of reasons as such but in the existence of *weighty* reasons, or at least in reasons that are relatively weighty in their situation.[13] This properly turns our attention to the question of what explains the weights of reasons. The value-based approach seems to have the right materials to provide such an account, since it is clear that in many cases at least the weight of reasons varies, other things equal, with the value of the state of affairs that would be promoted by the favored action.[14] Let's start with a stronger such principle, one according to which weight is *fully* explained by the value of the relevant state of affairs:

NAÏVE WEIGHT: The weight of a reason of the form [*x*-ing would promote S] is a monotonically increasing function *just* of the expected value of S.[15]

Our puzzle suggests that we need to reject NAÏVE WEIGHT. Each negative reaction to the shooting of Pelé is valuable, and hence each generates a distinct reason to shoot Pelé.[16] The combined weight of these reasons—

[12] Following ordinary practice I will often use the terms "value" and "is a reason" in a way that is neutral with respect to their valence: i.e., such that negative value is value and reasons against are reasons. I use "reason not to *x*" and "reason against *x*-ing" interchangeably. I distinguish both from "a reason to prevent . . . ," which is a reason for a different action.

[13] Here I'm following a tradition including Raz (1976) and Schroeder (2007) according to which some reasons are so lightweight as not to be worth mentioning.

[14] For a fuller defense, see my (2016).

[15] The "expected" qualification is in there to avoid an irrelevant distraction. It might be helpful here to say a word about the normative significance of the agent's evidence, and in particular how that evidence affects the weights of reasons for action. A hardcore objectivist view would deny that evidence affects weight at all. A hardcore subjectivist view would insist that only evidence about the relevant actions, relations, and values—even if misleading—is relevant to weight. Middle-of-the-road views would allow that evidence can modify weight in various ways. I submit that the issues of particular interest to us in this paper—various other forms of modification and conditionality—are orthogonal to these more familiar issues about the relationship between evidence and weights of reasons for action. I'll say no more here about this variable. We will assume henceforth that the agent knows (or stands in the relevant evidentiary relation to) all the normatively relevant facts.

[16] If we separate the value of the intentional attitude from the disvalue of the experience, we get two reasons the net weight of which is still positive; we'll henceforth ignore this possibility.

assuming a non-exotic principle concerning the aggregation of weight—will be proportional to the overall value of the shooting plus all these reactions, and we've seen that this overall value is positive. This will get us the result that you have more reason to shoot Pelé than to not. This is, again, intuitively false. We must therefore reject NAÏVE WEIGHT on the same grounds that we rejected Consequentialism.

We still have the following weaker principle available, namely:

SIMPLE WEIGHT: The weight of a reason of the form [*x*-ing would promote S] is a monotonically increasing function of the expected value of S, other things being equal.

This principle allows that *other* considerations might affect the weight of some reason, besides the expected value of the relevant state of affairs, S. We accept SIMPLE WEIGHT, and hence require any subsequent principle to be consistent with it. But SIMPLE WEIGHT is not very theoretically satisfying on its own. Ideally we want a systematic explanation for the weight of reasons that also applies to cases like Pelé's. Indeed it is our hope that by attending to Pelé's case we will uncover a systematic principle that applies to normal cases, and to cases like Pelé's, and to a host of other cases involving extrinsic value.[17] So we turn now to consider more sophisticated principles consistent with SIMPLE WEIGHT.

SECTION FOUR: A FIRST PASS AT A SOLUTION

I want to suggest, as a first pass, that if a state of affairs (the *object* state of affairs) has some value only in virtue of the fact that a certain relation obtains to some other obtaining state of affairs (call that whole fact the *condition*), then the value of the object state of affairs generates a reason only if the whole consisting of both is valuable. This is compatible with having *other* reasons to promote the object state of affairs, since the object state of affairs may also have intrinsic value or some other extrinsic value.

[17] Importantly, I don't mean to imply that this will amount to a full theory of weight, since there may well be other kinds of modification—temporal discounting, intensification arising from personal commitments, attenuation arising from one's engagement in value-realizing institutions, etc. Let's continue to ignore these.

Specifically we'll start with the following principle:

CONDITIONAL REASONS: The fact that your *x*-ing would promote S1 is a reason with non-zero weight for you to x if and only if: (1) S1 is finally valuable, and (2) if the fact that S1 is finally valuable obtains in virtue of some relation between S1 and some other state of affairs S2, then the whole consisting of S1, S2, and the relation between them is valuable overall.

This principle captures the intuitive idea that whether we have a reason to promote some valuable state of affairs depends upon whether and why it would be valuable. When the relevant state of affairs would be valuable only because of some relation to some other disvaluable state of affairs, then plausibly your reason to bring about the first valuable state of affairs depends in some way on whether it would be valuable overall if the whole consisting of both obtained.[18]

We are assuming here that the value of a reaction to the disvaluable state of affairs is not absolutely more than the disvalue of the state of affairs to which the individual is reacting. Since we are assuming away any *other* values, it follows that the value of any whole consisting in Pelé's shooting plus one negative reaction will be negative. Given CONDITIONAL REASONS, you would have no reason to shoot Pelé arising from the value of the reaction of any one individual. Consequently, it seems you have no reason to shoot Pelé.

Importantly, this is a *familiar* axiological relationship. This provides some intuitive support for the explanatory priority of value, and for the appeal to axiological distinctions in order to explain this deontic phenomenon.

SECTION FIVE: BASIC VALUES AND OVERLAPPING REASONS

This is promising, but there is more work to be done.[19] Grant that each whole consisting in Pelé's death and one sad reaction is disvaluable overall, and you would have no (non-zero weighted) reason to bring it about. But we have already suggested that the overall state of affairs

[18] I don't assume that S1 and S2 are distinct. When S1 is S2, CONDITIONAL REASONS maintains that the fact that *x*-ing would promote S1 is a reason for you to *x* if and only if S1 is finally valuable. In other words, CONDITIONAL REASONS implies VALUE-REASONS for the special case in which S1 is identical to S2, i.e., when S1 is intrinsically finally valuable.

[19] This section applies the theory of overlap developed in more detail in Maguire (2016).

consisting in Pelé's death and millions of sad reactions would be positively valuable overall. So why doesn't *this* state of affairs provide you with a reason to shoot? To answer this, we need some account of when reasons *overlap*.

Suppose that by pressing some button once you would make it the case that two dogs, in separate rooms, would each enjoy a delicious treat. If you fail to press the button, nothing disvaluable will happen, and nothing valuable would come of the delicious treats. Plausibly the happiness of each dog would give you a reason to press the button. You have two reasons, we can say, to press the button. But do you have a third reason, arising from the state of affairs consisting in the two dogs being satisfied, in addition to each reason arising from one of the dogs' being satisfied? Clearly not. Perhaps we can *say* that this state of affairs constitutes a reason to press the button. But this wouldn't count in favor of pressing the button in addition to the two reasons respectively provided by the each of the dogs being separately satisfied. Neither would the fact that an intelligent sentient creature (a more abstract state of affairs) would be satisfied provide you with an additional reason to be added to the reason provided by the fact that one dog will be satisfied.

To make this a little more precise we can appeal to the distinction, familiar to axiologists, between *basically* and *non-basically* valuable states of affairs. Intuitively, the basically valuable states of affairs are the states of affairs that are the *source* of an evaluative contribution (whether positive or negative) to the world on their own.[20]

For the sake of definiteness, I'll offer a version of Fred Feldman's (2000) seminal account:

BASIC VALUE: a state of affairs is *basically* valuable if and only if it is a pure attribution of a finally valuable property or relation.

One's substantive axiology will specify which properties are such that pure attributions of that property to an object are finally valuable (and under what conditions, if any). For an attribution of a property to be *pure* is for it to refer directly to an object (i.e., not by description), and to

[20] I say they are the "source" of an evaluative difference rather than that they themselves make an evaluative difference, because some basically valuable states of affairs are *conditionally* valuable. (On the source/condition distinction, see Bader forthcoming.) The basic/non-basic distinction crosscuts the intrinsic/extrinsic distinction.

do so unrepeatably.[21] Many valuable states of affairs fail to meet these conditions, usually by including insufficient or superfluous information.

Some states of affairs are not basically valuable because they contain insufficient information—they are too abstract. The state of affairs of [at least one person's being happy to degree *n* for duration *d*] is not basically valuable, since it doesn't contain enough evaluatively relevant detail; specifically it is not a pure attribution. This state of affairs obtains in virtue of some more fundamental state of affairs, such as [Jones is happy to degree *n* for duration *d*]. Indeed, this state of affairs might be metaphysically overdetermined by many different people's happiness. For the same reason, the state of affairs of [two utterly unconnected dogs being happy to degree *n* for duration *d*] is not basically valuable.

Some valuable states of affairs are not basically valuable because they include irrelevant information. For instance, the state of affairs of [one person's being happy to degree *n* for duration *d* within one million miles of *m* other planets] contains superfluous detail; the fact about planets is evaluatively irrelevant. The granularity of the basically valuable states of affairs will be specified by the appropriate axiological theory. Consider the old issue in axiology concerning whether sadistic happiness is valuable. The broader question here is whether the evaluative valence of the intentional object of a state of happiness is relevant to the evaluation of that state. Suppose that according to our axiology it is not. Then such states of affairs as the following would not be basically valuable: [Jones is happy to degree *n* for duration *d* from gazing at the stars]. On this axiological theory, the bit about the stars is evaluatively superfluous.

Then we can enter the following principle:

OVERLAP: Two reasons *overlap* just in case they are explained by the same basically valuable state of affairs.

The reason to make two dogs happy overlaps with the reason to make one dog happy and the reason to make the other dog happy. Neither of these reasons overlaps with the reason to make the dogs' owners happy, or the reason to make yourself happy by seeing the happy dogs. The

[21] In a full dress version we would add a spatio/temporal/worldly index to ensure unrepeatability. See Feldman 2000.

reason to make Mildred happy overlaps with the reason to make *someone* happy, when that someone is Mildred.

Now to reply to the objection. The value of the overall state of affairs consisting in Pelé's death and the millions of sadnesses is not basically valuable. The basic values in play are the disvalue of Pelé's death and the values of each of the sadnesses. Now suppose that there were some reason arising from the whole consisting in all of the responses to Pelé's death. This whole is not a basically valuable state of affairs. It has no *additional* organic value.[22] Its aggregate value is just the sum of the values of its parts. Any reason to promote the whole would overlap with any reasons to promote its parts. But there are no reasons to promote these parts. Consequently, there is no reason to promote this whole.

Here's another way to put the point. Facts about actions realizing states of affairs that are non-basically valuable—e.g., the fact that pressing the button will make two dogs happy—do not contribute weight to the action in question *on their own*. Only facts about actions realizing states of affairs that are basically valuable contribute weight to the action in question. A variety of other facts, including facts about the realization of non-basic value, might well be cited as reasons in reason-giving contexts (deliberation, advice, justification), and might as well be called non-basic reasons. The fact that shooting Pelé would promote the large aggregate state of affairs is a fact about an action realizing a non-basically valuable state of affairs. It overlaps with reasons to bring about the virtuous responses and the reasons not to shoot Pelé. It does not contribute separate weight to the shooting of Pelé. These reasons do not affect the overall weight of reason to shoot Pelé. Therefore, there is no reason to shoot Pelé.

If we replace CONSEQUENTIALISM with CONDITIONAL REASONS, and throw in OVERLAP, then we can accept all of our earlier assumptions, while still denying that you have reason to shoot Pelé. Once we add a principle relating reason and ought, we can also deny that you ought to shoot Pelé. Similar reasoning will apply to your tripping up of pensioners, kicking of dogs, etc.

[22] Of course, there are other ways of describing the situation in which it does, e.g., if solidarity is promoted amongst the many sufferers. We're assuming this away under NOTHING ELSE AFOOT.

SECTION SIX: INTRINSIC VALUE AND RESTRICTED SEPARABILITY

It may help to draw out the deontic significance of this pair of principles (OVERLAP and CONDITIONAL REASONS) by comparing our original Pelé case with the simplest kind of case for a value-based theory of reasons. In the simplest cases, some action available to you would cause some further states of affairs.[23] These states of affairs may have a range of values: some may have no intrinsic value; others will have intrinsic value or disvalue. In such simple cases we exclude all kinds of extrinsic value. Now let's consider a specific simple case, one in which the action will promote some disvaluable state of affairs, but also cause a number of distinct intrinsically valuable states of affairs. Here's an approximate[24] example of such a case:

GOING TO THE MALL: You dislike going to the mall. It is noisy and full of people trying to sell you fashionable scarves. There are various reasons you might have to go to the mall: to eat a fancy burger, to buy a scarf, or to find an unusual present for your uncle Albert. The unpleasantness of the mall is more absolutely disvaluable than any one of these things is valuable. Alas! Today you need all three of these things. These reasons aggregate and together outweigh the reason not to go to the mall provided by its unpleasantness.

It is quite plausible that each valuable or disvaluable state of affairs in that set of causal consequences constitutes a reason for or against performing the action, and that the weight of that reason is a function of the value of that state of affairs. Our NAÏVE WEIGHT principle, that the weight of a reason is a monotonically increasing function just of the value of the relevant state of affairs, is most plausible in these simple cases. The values of going to the mall yield reasons that add up separately, and will eventually outweigh the reasons against provided by the unpleasantness of mall-shopping. Such cases provide intuitive motivation for the following thesis:

RESTRICTED SEPARABILITY: reasons explained by distinct intrinsically basically valuable states of affairs contribute their weight separately.

[23] For simplicity let's allow ourselves to talk about causal relations between states of affairs.

[24] Only approximate since the putative intrinsic values certainly aren't basically valuable. Also for simplicity we ignore the difference between the presence of value and the absence of disvalue.

It is an advantage of a value-first account of reasons that it yields this RESTRICTED SEPARABILITY thesis. But it would be a mistake to generalize from such simple cases to exceptionless principles governing weight or separability. Matters are more complicated when the values of states of affairs among the consequences depend in various ways upon the values of other states of affairs in the consequences. It remains plausible that the weights of reasons to promote extrinsically valuable states of affairs depend upon facts about the conditions of these extrinsic values.

This contrast between the deontic significance of intrinsic values and extrinsic values gives us a reply to another objection.

Suppose that the great outpouring of sadness after Pelé's death would create a strong sense of solidarity and community across Brasilians and football fans around the world. This would be an additional value instantiated precisely by the whole consisting in the many sadnesses, and not located aggregately at each particular sadness. Given OVERLAP this would generate another reason to shoot Pelé, and if the value of the solidarity is great enough, perhaps this reason would even outweigh the reason not to shoot. Indeed, we can go on, and imagine that this great communal grief brings peace to warring sets of football fans or inspires young Brasilian children to improve their own lives. Wouldn't these facts generate various (non-zero weighted) reasons precisely to shoot Pelé?

Indeed they would. But there is no objection here to our puzzle or to CONDITIONAL REASONS. The many values instantiated by the many improvements in children's lives are intrinsic values, not extrinsic values. Each fact of the form [shooting Pelé would promote an improvement in child *a*'s life] constitutes a reason to shoot, the weight of which is a function of the intrinsic value of the improvement in *a*'s life. For all I have said so far,[25] enough of these will eventually outweigh the reason not to shoot Pelé provided by the disvalue of his death. Some deontological views maintain that no amount of promoted value or no amount of avoided disvalue could justify murder. Such views will insist that murder is always prohibited. I reject this position. (But even if you don't, bear in mind that I'm not resting my case on the significance of *murder* as such—so you are free to pick some other significant disvalue

[25] Of course, a full ethical theory would need to say more, for instance about the ethical significance of the doing/allowing or intending/foreseeing distinctions. See Wedgwood 2009 for an account in terms of value and modification.

that you think can be outweighed.) Given the possibility that there can be reasons to murder, I see no obstacle to allowing—for instance—that the many small improvements in the lives of the young Brasilian boys and girls constitute so many fairly lightweight reasons to shoot Pelé. Then it is just an open question how many improvements give you more reason to shoot than not—presumably a very large number. The crucial point is that the disvalue of the shooting doesn't modify and hence doesn't disable the weight of each of these reasons to shoot.

The values of the virtuous sad responses do not generate weighted reasons to shoot, but the values of the small improvements in the lives of children do. To put the point more abstractly: reasons are defeated when one's action would promote the condition on the extrinsic value of an outcome, but not when one's action would promote a disvaluable means to an intrinsically valuable outcome. Is the distinction between intrinsic and extrinsic value sufficient to explain such a difference in deontic significance?

SECTION SEVEN: EXTRINSIC VALUE AND NON-SEPARABILITY

Let me provide more motivation for the claim that this difference in deontic significance is explained by the distinction between intrinsic and extrinsic value. We start with two observations. First, we are familiar with a kind of interdependence between values. Even if we reject Hurka's principles of virtue for being too strong, we will accept that there are some cases in which the value of something depends either upon the obtaining or upon the value of something else. Second, the non-separability of reasons is familiar in the literature on reasons. I would even say this is the default view. So it is certainly not a problem for a theory of reasons that it entails some restricted non-separability. On the contrary, the fact that a theory of reasons yields a thesis like RESTRICTED SEPARABILITY is a consideration in its favor.

Now think about the underlying rationale for the value-based theory of reasons. Reasons are facts about which actions would promote valuable states of affairs. So take any pair of an action and a state of affairs such that, although the state of affairs would be finally valuable, the whole consisting of the action and that state of affairs necessarily would be disvaluable. It is plausible, given this underlying rationale, that there

would be no reason (or at most a reason of zero weight), to perform that action for that reason.

Now think about the difference between the Pelé case and GOING TO THE MALL. In both cases there are states of affairs that are valuable promoted by the respective actions. But in GOING TO THE MALL the relation between the action and the outcomes is merely contingent. It just so happens that going to the mall is the only way to get new shoes and it just so happens that going to the mall is unpleasant. The relevant object states are *axiologically distinct*.

Perhaps you are tempted to resist this conclusion on the grounds that the action is *instrumentally valuable*. It is instrumentally valuable in virtue of standing in the relevant causal relation to the relevant outcome. Hence these states of affairs are not axiologically distinct. It is important to see what is wrong with this thought. We have so far said nothing whatsoever about instrumental value in connection with GOING TO THE MALL. We have just been discussing intrinsic values—the disvalue of the unpleasantness, the intrinsic value of eating the burger, etc. In effect, the instrumentality is built into the regimentation of value-based reasons, for they are essentially facts of the form [doing something *is instrumental to* some state of affairs].[26] All reasons are instrumental reasons. Reasons consisting of the fact that some action will itself instantiate some property are just special cases in which the promotion relation is mere instantiation. I'm content (for now) not to discuss instrumental value at all. The difference between GOING TO THE MALL and Pelé is explained by the difference in the deontic significance of intrinsic value and extrinsic value, not by the difference between instrumental value and either of these.

So when states of affairs are axiologically distinct, we have a default case for RESTRICTED SEPARABILITY. When states of affairs are not axiologically distinct, their deontic significance is not separable.[27] The fact that

[26] I assume that instantiating value is the limiting case of being instrumental to it. Of course, it is often felicitous to appeal to facts not regimented in this canonical form as reasons for this or that: e.g., the fact that the pubs close in twenty minutes is a reason to hurry up. Elsewhere I argue that all reasons are explained in terms of reasons regimented as in the main text, in [action-promotion-object state] form (2016).

[27] There are two ways in which states of affairs might fail to be axiologically distinct, either because the value of one depends upon the obtaining of the other, or because the value of one depends upon the value of (and the obtaining of) the other. I assume that CONDITIONAL REASONS applies to both.

the death leads to the sadnesses makes the sadnesses extrinsically valuable. But this value will be essentially part of a whole that is negatively valuable overall. Moreover this is not merely contingent; it is a substantive axiological thesis. This result is entailed by CONDITIONAL REASONS.

SECTION EIGHT: CONDITIONALITY AND MODIFICATION

But CONDITIONAL REASONS won't quite do the trick. For even though the state of affairs consisting in Pelé's death and one sad reaction is less valuable than neither, once Pelé is dead (perhaps some less enlightened Consequentialist shot him), it would be better for there to be one sad reaction than none, and clearly our reasons track this. But the value of the whole won't change. So if whether you have a reason to bring about sadness in response to Pelé's death, say by telling someone about the shooting, depends upon whether the whole consisting of the death and the sadness is valuable, then clearly you won't have a reason to do so. But this now seems like the wrong result. Once the deed is done, the reasons change.

We will see this point more clearly if we switch examples. Let's talk about this in terms of retributive punishments—simply assuming that retributive punishment is valuable.[28] Suppose that some crime involving direct and considerable harm has been committed. Let's suppose that the appropriate punishment for some crime is ten years in jail. Presumably, and in line with Moore's thesis discussed in section 2.2, the value of the punishment would be less than the absolute value of the crime, such that the overall value of both is negative, i.e., disvaluable. However, once the crime has been performed, things will be better if the criminal is punished than not.

So far this doesn't seem to present a principled difficulty for the project of explaining reasons in terms of values. It does present a problem for CONDITIONAL REASONS since the whole consisting of the crime and the

[28] The structure of this case is the same but we avoid two difficulties specifically concerning attitudes, namely whether there are any reasons for attitudes at all, and whether we have reasons to bring about attitudes in other people. Unfortunately, we encounter a new complication, concerning how best to account for the reasons pertaining to institutional roles, e.g., being a judge or sheriff. Since I'm presenting this case as a problem for my view I'm not too worried about motivating these cases further. Nothing I say in this paper turns on the particular substantive examples. It doesn't matter whether you believe that retributive punishment is valuable. What matters is whether, if so, our account of the weight of reasons delivers the intuitively correct results. Also the law/morality distinction is not relevant to the point.

punishment would be disvaluable whether or not the crime has been committed. Hence if any reason to punish depends just upon whether the whole consisting in the crime and the punishment is positively valuable, there would be no reason to punish (or at best a reason with no weight). We need to add something to the principle that will be sensitive to the status of the relevant condition.

Let's go back to the rationale. The deepest value-first thought is that deontic facts—in the first place, *reasons*—are explained by the values we can *affect* by our actions. Once the crime has been committed, the damage has been done. It is not possible (given the current state of technology) to make it the case that this crime did not occur. The crime involved a significant harm. In this case the disvalue of this harm can also not be affected by anything you can now do.

Interestingly, this isn't always the case. We do sometimes have reasons to do things to make past states of affairs better, by realizing a condition on their value: e.g., to make some effort constitutes an achievement. One might think about the value of retribution on this model, as attenuating the disvalue of the crime. However, this model strikes me as less intuitive here, at least for some crimes. But it doesn't matter for our purposes which model we adopt, for your options are either to make one thing less disvaluable or to bring about value. So long as reasons are explained by values you can affect, either of these will yield a reason to punish.

In short, once the condition on some extrinsically valuable state of affairs already obtains, that state of affairs has the same deontic significance as an intrinsically valuable state of affairs. So we need to restrict the conditionality of CONDITIONAL REASONS to conditions that do not already obtain. One simple way to do this is to route the explanation through reasons to promote the condition. This is simple because it is usually assumed that there are no reasons (or at most reasons of zero weight) to promote states of affairs that already obtain. Thus:

MODIFYING REASONS: The fact that your *x*-ing would promote S1 is a reason with non-zero weight for you to *x* if: (1) S1 is valuable, (2) the weight of the reason is an increasing function of the value of S1, and (3) if the fact that S1 is valuable obtains in virtue of some other state of affairs S2 (where this might include some relation to S1), then the weight of your reason to promote S1 is modified by the weight of your reason to promote S2.

Again we assume that the same holds, *mutatis mutandis*, for disvalue and reasons against. In order to work out the weight of the reason not to S2, we apply MODIFYING REASONS again. For intrinsically valuable states of affairs, MODIFYING REASONS (like CONDITIONAL REASONS) reduces to VALUE-REASONS, i.e., reasons are explained directly by intrinsic values and in proportion to their weight, other things being equal. Hence when S2 is intrinsically disvaluable, as in Pelé's case, the reason against shooting will be proportional to the disvalue of the death. S2 is disvaluable and S1 is valuable. So long as Moore's thesis holds in the case, S2 is guaranteed to be absolutely greater than S1. Hence the weight of S1 will be attenuated either down to zero, or even to a negative amount. We'll assume that the valence of reasons cannot change.[29] Hence given MODIFYING REASONS and MOORE'S THESIS the weight of any reason to shoot Pelé would be zero.

But in a case in which the condition already obtains—someone has committed a crime, and there is a question whether to punish proportionately—the weight of the reason to punish will be a function of the (extrinsic) value of the punishment. There is no reason now not to commit the crime or aid in the committing of the crime, just like you have no reason to shoot JFK. Consequently, there is no attenuation by such a reason. The weight of the reason is just a function of the extrinsic value of the punishment (modulo irrelevant sources of modification).

There is a further important difference between MODIFYING REASONS and CONDITIONAL REASONS, namely that MODIFYING REASONS supplies a modifier for the weight of the relevant reason rather than a condition on its having non-zero weight. A modifier attenuates or intensifies the weight of a reason. On the assumption that modification cannot change the valence of a reason, and a simple summative model of modification, a modifier with greater negative weight than the positive weight it modifies will reduce that weight to zero. This is what happens in the Pelé case. In other words, this difference between conditionality and modification is practically irrelevant in cases like this. MODIFYING REASONS entails CONDITIONAL REASONS in our original Pelé puzzle.

[29] The rationale for this is that the value of the state of affairs realization of which explains the *existence* of the reason, rather than any values that affects its *weight*, explains its valence.

CONCLUSION

We are left with a rather powerful principle governing the transmission of weight. This principle applies to all cases of extrinsic value: a rather wide-ranging brief. In addition to cases like sadness at tragedy, we have various kinds of part value (synchronic, diachronic, interpersonal, intrapersonal), symbolic value, and perhaps instrumental value. Indeed, if Samuel Scheffler (2013) is right that the significance of much of what we care about is conditional upon the persistence of humanity for a while, then much more of what we take to be intrinsically valuable will turn out to be extrinsically valuable. It could be that most value is extrinsic value. It is for future work to explore the implications of principles like MODIFYING REASONS to these cases. My main goals in this paper are more limited.

The three goals of this paper were to draw out in detail this tension between prominent value-based theories of action and virtue, to offer a solution that covers the cases and proffers further interesting and useful upshots, and ultimately to show that non-Consequentialist value-first ethical theories have a wide variety of machinery at their disposal in pursuit of an ethical theory sensitive to pre-theoretic intuitions.[30]

REFERENCES

Bader, R. (Forthcoming). "Conditions, Modifiers, and Holism," in *Weighing Reasons*, ed. E. Lord and B. Maguire. Oxford: Oxford University Press.

Bader, R. (ms). "Two Value-Theoretic Trichotomies."

Feldman, F. (2000). "Basic Intrinsic Value," *Philosophical Studies* 99: 319–46.

Hurka, T. (2000). *Virtue, Vice, and Value*. Oxford: Oxford University Press.

Maguire, B. (2016). "The Value-Based Theory of Reasons." In *Ergo*.

Maguire, B. and Lord, E. (2016). "An Opinionated Guide to the Weight of Reasons," in *Weighing Reasons*, ed. E. Lord and B. Maguire. Oxford: Oxford University Press.

Moore, G. E. (1903). *Principia Ethica*. London: Cambridge University Press.

Raz, J. (1976). *Practical Reason and Norms*. Oxford: Oxford University Press.

Scanlon, T.M. 1998. *What We Owe To Each Other*. Cambridge: Havard University Press.

Scheffler, S. (2013). *Death and the Afterlife*. Oxford: Oxford University Press.

Schroeder, M. (2007). *Slaves of the Passions*. Oxford: Oxford University Press.

Wedgwood, R. (2009). "Intrinsic Values and Reasons for Action," *Philosophical Issues* 19: 342–63.

Zimmerman, M. (2001). *The Nature of Intrinsic Value*. London: Rowman & Littlefield.

[30] Many thanks to Julia Driver, Daniel Fogal, Tom Hurka, Eden Lin, Errol Lord, Doug Portmore, Geoff Sayre-McCord, Karl Schafer, Sarah Stroud, David Velleman, Jack Woods, and audiences at NYU, UNC Chapel Hill, and the Arizona Workshop in Normative Ethics.