Synthese (2024) 204:135
https://doi.org/10.1007/s11229-024-04751-4

ORIGINAL RESEARCH

The co-evolution of virtue and desert: debunking intuitions about intrinsic value

Isaac Wiegman[1] · Michael T. Dale[2]

Received: 11 January 2024 / Accepted: 16 August 2024

Abstract
Thomas Hurka's recursive account of value appeals to certain intuitions to expand the class of intrinsic values, placing concepts of virtue and desert within the realm of second and third order intrinsic goods, respectively. This is a formalization of a tradition of thought extending back to Aristotle and Kant via the British moralists, G. E. Moore, and W. D. Ross. However, the evidential status of such intuitions vis a vis the real, intrinsic value of virtue and desert is hostage to alternative explanations. If there is a plausible competing explanation for these intuitions, then the (putative) fact that desert and virtue are intrinsic (rather than instrumental or derivative) goods seems a much less obvious choice for the best explanation. As it turns out, there are plausible evolutionary explanations for these intuitions about desert and virtue. These evolutionary explanations suggest that it is adaptive to value desert and virtue separately from their instrumentality for other goods. Consequently, these explanations debunk intuitions about the intrinsic value of desert and virtue.

Keywords Intrinsic value · Value realism · Indirect reciprocity · Virtue · Desert · Evolutionary debunking · Evolutionary psychology · Evolutionary game theory · Evolution of cooperation

✉ Isaac Wiegman
isaac.wiegman@txstate.edu

Michael T. Dale
mdale@hsc.edu

[1] Texas State University, San Marcos, TX, USA

[2] Hampden-Sydney College, Hampden Sydney, VA, USA

Published online: 09 October 2024

1 Introduction

It is important to know what intrinsic goods there are, if any. Thomas Hurka's recursive account of objective value aims to considerably expand the class of intrinsic goods (e.g., pleasure, knowledge, satisfaction of preferences) by including virtue and desert (Hurka, 1992, 2001). Moreover, this account captures several key intuitions about, and symmetries between, virtue and desert that have been influential in moral theories from Aristotle to Kant to Moore and Ross. Hurka argues that these symmetries are an additional reason to think that desert and virtue are intrinsic goods. As interesting and compelling as these intuitions and symmetries are, their evidential value vis a vis intrinsic goodness is hostage to alternative explanations. If there is a plausible competing explanation for these symmetries and intuitions, then the (putative) fact that desert and virtue are intrinsic goods looks a much less obvious choice for the best explanation. In this paper, we argue that there are plausible evolutionary explanations for the symmetries that Hurka identifies and perhaps also the intuitions on which these symmetries are based. These evolutionary explanations suggest that it is adaptive to value desert and virtue separately from their instrumentality for other goods and to value them in accordance with the symmetrical constraints that Hurka identifies. If these explanations work, then our tendencies to value virtue and desert as intrinsic goods are perhaps better explained by their derivative value (i.e., the fact that they have certain advantages or bring about certain outcomes) for various selection problems. This shifts the burden of proof in favor of the hypothesis that the value of virtue and desert is derived from their effects.

In the following section, we will lay out Hurka's recursive theory of intrinsic goods and one particular symmetry that it establishes between virtue and desert. In section three, we offer an evolutionary explanation of the co-evolution of virtue and desert that explains why they would have this symmetry. In section four, we argue that this explanation debunks the intuitions and symmetry as evidence for a realist account of the intrinsic value of virtue and desert.

2 Hurka's recursive theory of Intrinsic Goods: intuitions and symmetries

Given some list of first-order goods that are intrinsically valuable (e.g., pleasure, knowledge, or difficult achievements), Hurka points out that we can expand the list by adding a recursion clause stating that: "If x is intrinsically good, loving x for itself is also intrinsically good" (1992: 150). Given a list of intrinsic evils (e.g., pain, false belief, or failure), he adds that it is intrinsically good to hate what is evil. Moreover, if we say that virtue is loving the good (i.e., "desiring, pursuing, or taking pleasure in it") for itself and hating what is evil (i.e., "desiring or pursuing x's nonexistence or being pained by its existence") for itself, then we can say that virtue is intrinsically good. If we add that vice is loving what is evil and hating what is good, we can add further recursion clauses that identify vice as intrinsically evil.

In addition to these second-order intrinsic goods (and evils), Hurka adds plausible principles involving desert. For example: "The combination of virtue and pleasure [or vice and pain] in the same person's life is intrinsically good as a combination…" (2001: 10). Here, Hurka draws on a tradition that dates back as far back as Aristotle (1984: 1175b24–30) and Kant, the latter of which claimed that the highest good occurs when happiness is "in exact proportion with the morality of the rational beings who are thereby rendered worthy of it" (Kant, 1899: 456). Hurka includes principles that cover each combination of virtue, vice, pleasure and pain (i.e., concerning the combinations of virtue and pain, vice and pain, vice and pleasure) and thereby capture intuitions about desert (e.g., that the combination of pain and vice is intrinsically good as a combination). So, desert goods (and evils) are intrinsically good (or evil) as a combination of first and second-order goods. This theory is supposed to capture much of what Moore (1993) discussed under the heading of "organic wholes"; that higher-order values (e.g., the good of desert) arise that are more than the combination of lower-order parts (i.e., the pleasure being experienced and the moral virtue of the person experiencing it). Moreover, it provides a way of unifying some of Ross's fundamental axiological commitments:

> Four things, then, seem to be intrinsically good – virtue, pleasure, the allocation of pleasure to the virtuous, and knowledge (and in a less degree right opinion). And I am unable to discover anything that is intrinsically good, which is not either one of these or a combination of two or more of them (2003: 140).

Hurka lists several other attractive features of the recursive account of virtue, of which we mention only two. First, it explains certain intuitions about the total value of possible worlds where a non-recursive theory of first-order goods (like hedonism) cannot. To illustrate this, Hurka asks us to imagine two possible worlds:

> In the first world, natural conditions are benign, and, because of this, people live very pleasantly. But they are all self-concerned: they do not desire or pursue each other's pleasure, and their hearts are cold to any enjoyments not their own. In the second world, resources are more scarce and there is some disease, but people are more altruistic. They care about each other's happiness and are pleased when it is attained. This altruism increases their happiness, but not enough to overcome their natural disadvantages. Despite their mutual concern, people in the second world enjoy less pleasure overall than those in the first (1992: 154).

On a hedonistic theory of value—one that recognizes only happiness as a good—the first world is clearly the better one. Nevertheless, if we had the option of creating only one of these worlds, it seems many reasonable people would choose the second. This thought experiment is similar to Ross's "two worlds" argument, which concerns not only the value of virtues like altruism (etc.) but also the value of desert:

> If we compare two imaginary states of the universe, alike in the total amounts of virtue and vice and of pleasure and pain present in the two, but in one of which the virtuous were all happy and the vicious miserable, while in the other the virtuous were miserable and the vicious happy, very few people would hesi-

> tate to say that the first was a much better state of the universe than the second. It would seem then that, besides virtue and pleasure, we must recognize, as a third independent good, the apportionment of pleasure and pain to the virtuous and the vicious respectively (2003: 138).

Hurka's recursive account of value is consistent with our preferences across these examples, whereas hedonism is not. These kinds of examples offer some support not only for Hurka's theory but any theory that holds desert and virtue to be intrinsically good. What Hurka's recursive theory offers is a way of understanding these intrinsic goods in a way that is unified with our understanding of first order goods.

Hurka points out four different symmetries between desert and virtue, which suggest they are unified values. One of these symmetries is a direct implication of his recursive theory: that both virtue and desert are higher-level values that have specific relations to particular, lower-level values. And these relations are subject to a matching rule: "if an attitude's orientation matches the value of its intentional object… the attitude is appropriate or good" (2001: 12). For example, having a pro-attitude toward pleasure (one of the lower-level values) is good. Similarly, if the attitude does not match the value of its intentional object (loving what is bad), then the attitude is intrinsically evil. Since these relations are recursive, both the higher-level values and the responses to those values are intrinsically valuable.

Importantly, Hurka takes himself to be contributing to an objective theory of intrinsic goods. We can see this where Hurka lists his reasons "for exploring these parallels" (2001: 8):

> First, I find them intriguing in themselves. Philosophers have combined virtue and desert in the same moral theories since at least Aristotle without noticing the formal similarities between them or noticing them in all their detail. Second, the structure the two values share has an impressive internal integrity. For each of virtue and desert there are plausible conditions on its further specification that look independent, so we can wonder whether they are even compatible. But the conditions turn out not only to be consistent but in some cases to entail each other, so satisfying one requires satisfying the other. When a proposed value's structure has this internal coherence, this strengthens its claim to be a genuine value. Finally, the parallels themselves strengthen the claims of virtue and desert to be genuine values. A common objection to objective versions of consequentialism is that they contain only a list of intrinsic goods with no unifying explanation of what makes them goods. This objection is unreasonable if it assumes that only monistic theories of value are acceptable but not if it claims merely that, other things equal, a theory is more credible the more it can unify its various goods. But then the parallels between virtue and desert strengthen the case for a theory containing these values, just because they reveal a formal unity between them. (2001: 8)

First, Hurka's focus is on how our understanding of intrinsic values contributes to "objective versions of consequentialism." This qualifier would seem unnecessary if he were attempting to account for the intrinsic goodness of virtue and desert that remained metaphysically uncommitted to their objectivity.[1] Second, by saying that virtue and desert are "genuine" values, Hurka seems to be suggesting that they are objectively valuable. On a subjective reading of intrinsic value, what makes something valuable is just that the subject values it for its own sake. Having internal coherence or unity would seem unnecessary to prove that someone values virtue or desert for their own sake. So, it seems Hurka is claiming that these values are genuine as opposed to fake or "made up" by the individual. Unity and internal coherence are supposed to make it more plausible that virtue and desert exist independently of any individual's attitudes toward them and so, are not just subjectively imagined values.

So, Hurka's recursive theory captures several important, and seemingly objective, symmetries between desert and virtue, only one of which we tackle here: that they are subject to the matching rule regarding the valence of the attitude and the value of its object. To us, this symmetry does seem striking. Prima facie, it might seem plausible that desert and virtue are real, intrinsic goods and that we have some evidence for their status as such because of their common internal structure. We are not certain what Hurka's line of thought would be to support this evidential link, and it would seem to depend on assumptions about the moral psychology of value perception or acquisition, which may appear shaky under closer scrutiny. There could be a kind of poverty of the stimulus argument in the offing: If the only intrinsic goods were first-order goods (e.g., pleasure, knowledge, achievement) and if desert and virtue were merely good as instrumental for bringing about those goods, it is difficult to see why our grasp of their value would be so immediate, intuitive, unified, and coherent and difficult also to see why they would appear so early in development (as we discuss below). For example, the set of means to achieve a first-order good like pleasure are many and varied, even more so if we add to this set the means to knowledge or satisfaction or achievement. There is little reason to expect that individuals would come to directly value virtue and desert in symmetric ways if virtue and desert had to first be recognized as a (reliable?) means to pleasure (among other first-order goods). Nor do we seem to value these goods because we recognize them to be instrumental in bringing about first order goods. So, perhaps there is some faculty by which we recognize their value unmediated by their instrumentality for first-order goods. For these reasons, the hypothesis that virtue and desert are intrinsically good might seem much more plausible than the hypothesis that they are only good as instrumental for first-order goods.

Hurka's observations are not only striking from a philosophical perspective, where the theory choice is between normative theories but also from a psychological perspective, where the theory choice is much more expansive. How do people (individually and collectively) come to value desert and virtue in such similar ways? Evolutionary psychology may have a plausible answer. As it turns out, desert and virtue are derived from biological advantages and outcomes (e.g., stabilizing cooperation in a population), and their fittingness for these predicts substantial symmetry between the two. Even if Hurka's evidential claim (about the intrinsic and objective goodness of virtue and desert based on their unity and internal coherence) is suspect,

[1] Moreover, Ross and others that Hurka discusses in connection with the symmetries are explicitly realists about value.

the symmetries he notices are worth closer inspection and explanation. To that task we now turn.

3 Evolutionary models of reciprocity and the co-evolution of Virtue and Desert

The explanations we consider fall out of a family of evolutionary models of indirect reciprocity, which aim to explain certain patterns of cooperation in the animal kingdom. These models predict the co-evolution of virtue and desert, understood along similar lines as Hurka's recursive model. Models of indirect reciprocity have been elaborated relatively recently compared to other explanations of cooperation, and they also explain broader patterns of cooperation by comparison with these other explanations. For example, theories of kin selection explain some of the most widespread forms of cooperation in the animal kingdom (Hamilton, 1964), but they only predict cooperation among closely related individuals (i.e., ones likely to share the same genes). Nevertheless, there are broader phenomena of cooperation to be explained: widespread cooperation among unrelated individuals can be observed not only in humans but even between different species (as we describe below).

Theories of direct reciprocity explain only some of this cooperation: among individuals who interact frequently (Trivers, 1971). The original prisoner's dilemma is a well-defined game theoretic model that can be used to understand this latter kind of reciprocity.[2] The dynamics of the game change if repeated encounters are more likely. In such an iterated prisoner's dilemma, there is increased opportunity for trust-building, forgiveness, revenge, and so also, reciprocation. Famously, Axelrod (1984) took on the task of finding out what the best strategy would be in an iterated prisoner's dilemma (with computer programs). Here, a strategy defines how a player will act in each round of the game, depending (usually) on how the counterpart acted in the prior round. So, a typical strategy for this game will have an unconditional part (i.e., what to choose on the first n round), followed by a conditional part (i.e., what to choose if the counterpart defects after n rounds). For example, one of the best-known strategies is called Tit for Tat: cooperate on the first round and for each subsequent round match the counterpart's action from the prior round. If A follows a Tit for Tat strategy, then if B cooperates on the first round, A cooperates on the second round, and they both benefit from mutual cooperation. If, however, B defects, then A "takes revenge" on B by defecting back. This defection continues until B "asks for forgiveness" and cooperates again. When Axelrod put this strategy in competition with

[2] For those unfamiliar with this game theoretic scenario, imagine that two criminals, A and B, are arrested and individually questioned. Their options are the following. If A tells the police that B did the crime, A will be set free while B will serve ten years (and vice versa). If they both betray each other, each serves five years. But if they both remain silent, they will both only serve one year. Obviously, the best overall outcome (in terms of aggregate time served) would be for both of them to remain silent (i.e. mutual cooperation). However, defection is the safer choice, at least in this scenario, where there is only one iteration. From the perspective of A and B, each will get a better payout by defecting, regardless of which choice the counterpart makes: If B chooses to defect, then defection is better for A, because getting 5 years in prison is better than getting 10. If B chooses to cooperate, then defection is still better for A, because getting set free is better than one year in prison. This is why defection is the Nash equilibrium for this game.

various others, he found that Tit for Tat was one of the most consistently successful strategies. As such, Axelrod's model was initially taken as a way of explaining how cooperation could be stabilized through direct reciprocity, where individuals benefit directly from cooperation.

But as successful as the Tit for Tat strategy was in Axelrod's iterated prisoner's dilemma, it has its shortcomings. Most importantly for evolutionary models, it is not an Evolutionary Stable Strategy (ESS). An ESS is a strategy adopted within a population that cannot be invaded by alternative strategies. While Tit for Tat does a great job at beating out most other strategies, it cannot remain stable against invasion from all other strategies, even once it becomes fixed in a population. For example, Boyd and Lorberbaum (1987) showed that a population of Tit for Tat could be successfully invaded by a mixture of two strategies: Tit for Two Tats (which allows the other player to defect two times before defecting in return) and Suspicious Tit for Tat (which behaves like Tit for Tat, except that it begins with a defection).[3]

As interesting and valuable as these models are, they do not explain the full extent of cooperation in the animal kingdom. Specifically, they cannot explain cooperation where repeated interactions are unlikely or uncertain. Another limitation is that they seem incapable of capturing the discriminative capacities of any but the simplest of animals. For example, certain species of wrasse are known to clean the parasites off of reef fish in contexts where future interactions are uncertain (Bshary & D'Souza, 2005). This behavior is commonly understood as direct reciprocity, as both creatures directly benefit from the interaction. However, sometimes, instead of taking the time to find the parasites and unhealthy tissue on their "clients," the cleaner wrasse simply takes a bite of the reef fish's healthy tissue. This "cheating" behavior only occurs in a minority of interactions, but when it does occur, it usually results in the reef fish chasing the wrasse. Interestingly, if another reef fish observes this conflict, that fish is significantly less likely to offer themselves up for cleaning by the offending wrasse. Which is to say, reef fish *track* the actions of wrasse, and if they don't like what they see, they can decide not to use the wrasse's services. As illuminating as the theory of direct reciprocity is with regard to this cooperative arrangement, another, more complex model is needed to account for this type of reputation tracking. It enables individuals to benefit from cooperation (or pay the cost of defection) not only directly, that is, from interactions with one individual, but also indirectly, or because of their prior history of cooperation (or defection) with other individuals.

Game theoretic models of indirect reciprocity (IR) aim to explain these more discriminating forms of cooperation. As we have seen, even reef fish can implement a much more complex strategy in response to cleaner wrasses. That is, they are not just

[3] There are a few reasons why Tit for Tat is not an ESS, but one of them can be garnered from the fact that Tit for Two Tats gains an advantage over it in certain conditions. Tit for Tat is unable to deal with mistakes or defection on the first round (as with Suspicious Tit for Tat). If Tit for Tat is interacting with another player, and that other player defects for any reason (e.g., accidentally or out of "suspicion"), then Tit for Tat will defect on each subsequent iteration, and it will continue to defect as long as the other player defects. So, if the competing strategy pays back defection with defection (as does Suspicious Tit for Tat), then, after an initial round of defection, both strategies will be locked in a suboptimal cycle of defection. So, Tit for Tat will not secure the benefits of cooperation in the context of mistakes and "suspicion" as well as a strategy like Tit for Two Tats, which will sometimes cooperate in spite of defection.

programmed to cooperate on the first iteration and defect in response to defection. Rather, they are sensitive to factors that *predict* defection or cooperation, namely the prior history of a given cleaner wrasse. One might think that this kind of flexibility would be best explained by each reef fish acting rationally with an eye for its own best interest, but actually, their behavior can still be explained by evolutionary game theory (as in the case of direct reciprocity), where "it is not assumed that players are rational but only that successful strategies spread–by being inherited, for instance, or copied through imitation or learning" (Nowak & Sigmund, 2005, p. 1292). Like the prisoner's dilemma, these models involve two agents but unlike the prisoner's dilemma, only one agent gets a choice about what to do. This is sometimes called the *donation game*: When in the *donor* role, player A's options are to donate or not. If A chooses to donate, A will pay a cost to confer a greater benefit on A's counterpart, B. Whereas in the *receiver* role, A just gets a payout if their counterpart donates and nothing if not.

In computer simulations of this game, different players in a population are randomly paired with one another for a given number of interactions per generation. In certain conditions, cooperative strategies can evolve within simulated populations. In Nowak and Sigmund's early work on this (1998a; 1998b), cooperation can evolve if players are given *image scores*, which is an index of how cooperative an individual has been in the past. Players get a positive point added to their image score each time they donate in an interaction. With this in place, we can define the discriminating cooperator strategy in this way: they donate if and only if their counterpart has an image score greater than 0. However, the strategy is not stable against mutations. If mutations are introduced the strategy does not remain fixed in a population. If discriminating cooperators can have progeny with different donations strategies, then the population will continue to cycle from a majority of discriminating cooperators to unconditional cooperators, then to unconditional defectors, then back to discriminating cooperators.

Before we introduce further developments in these evolutionary models, it is worth mentioning that psychological research on reputation and reciprocity in humans is broadly consistent with these explanations of cooperation. Consider experiments on the donation game in humans. When the history of a person's giving (or not) was displayed to other participants, donations were significantly higher to those who had been generous to others than to those who had not (Wedekind & Milinski, 2000). Moreover, when people know that their reputation is at stake, they tend to be significantly more cooperative (Romano et al., 2017). These findings have been shown to be consistent across studies (van Apeldoorn & Schram, 2016; Levati & Greiner, 2005; Seinen & Schram, 2006) and cultures (Hu et al., 2019; Kato-Shimizu, Onishi, & Kanazawa, 2013). They indicate not only that people track the reputations of others (and themselves), but also that people's decisions about how much to cooperate with others are based on such image scoring to a significant extent.

This finding is further supported in the developmental literature. Meristo and Surian (2013) showed a group of 10-month-old infants events in which donors either fairly or unfairly distributed strawberries. When an unfair donor was subsequently rewarded, infants showed surprise (measured by longer looking times). When the fair donor was rewarded, infants did not show surprise. This indicates that infants at this age expect agents to conform to the principle of indirect reciprocity.[4] Similar results have been found among adolescents (Hu, Ma, & Luan, 2019), as well as among children in a less artificial setting (Kato-Shimizu, Onishi, & Kanazawa, 2013). All of this demonstrates that children from a very young age have the ability (and the inclination) to not only track image scores but also to make cooperative decisions based off of such scores. As these are no doubt complex tasks that require complicated cognitive processing, one of the most plausible explanations is that such reasoning is the result of innate cognitive mechanisms.

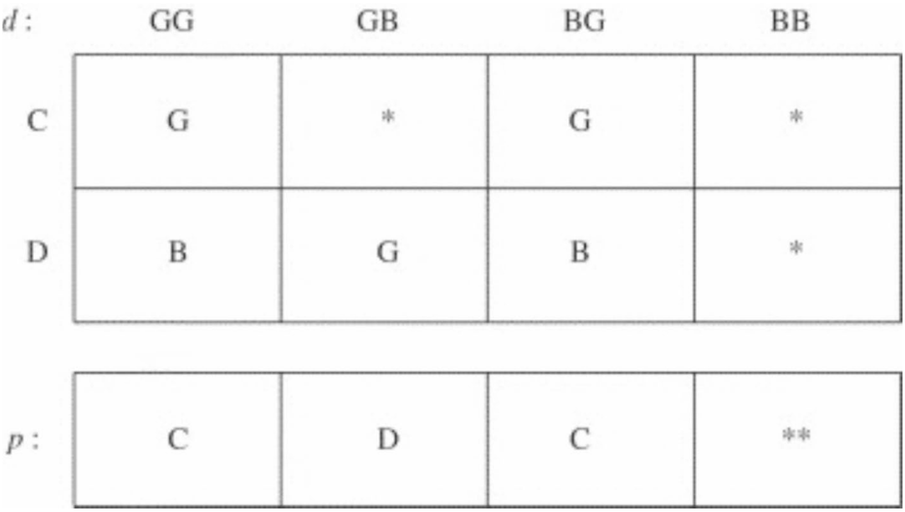
3.1 Indirect reciprocity and the matching rule

Returning now to Nowak and Sigmund's model, one problem with it is that image scores are straightforwardly determined by whether a player cooperates, regardless of whether they are cooperating with a known defector. So, unconditional cooperators sometimes get an edge on discriminating cooperators because the image score of the former never takes a hit from refusing to donate to a defector (or from mistakes in assessing the receiver's image score). The problem is that refusal to cooperate is exactly what makes discriminating cooperation resistant to exploitation by defectors.[5] So, a few years after Nowak and Sigmund introduced their model, Ohtsuki and Iwasa (2004) proposed a more nuanced set of options for how to calculate image scores, depending on the image score of the donor and the receiver. They evaluated a large set of possible "reputation dynamics," (*d*) which determine the reputation/image score of an agent depending on their action (cooperate or defect), their reputation (zero = bad, one = good) and on the reputation of their interaction partner. These reputation dynamics also determine whether an agent will cooperate or defect in a given interaction. Given these variables, there are 256 possible reputation dynamics (all possible variations on rows 1 and 2 in Fig. 1), and Ohtsuki and Iwasa (2004) used computer simulations to consider how each of these dynamics performed when paired with each of 16 possible behavioral strategies (all possible variations on row *p* of Fig. 1). Only eight of the reputation dynamics (depicted in rows 1 and 2 of Fig. 1), and two behavioral strategies (depicted on row 3 in Fig. 1) when paired with them, were able to outperform all 14 competing behavioral strategies. Basically, this is a way of measuring how each possible behavioral strategy performs under any possible socially shared system of norms for evaluating reputation. Ohtsuki and Iwasa (2006) proved analytically that each of these strategies is an ESS under a wide set of parameters and if any strategy is an ESS that maintains high levels of cooperation, it is one of the leading eight. In other words, these are the only social norms for evaluating reputation that will allow cooperation to be evolutionarily stable in a population on the basis of indirect reciprocity.

More recent research shows that four of the leading eight are more stable than the others under more plausible conditions. Ohtsuki and Iwasa were working with a simplified model where there were no errors in assessment, and where each individual's

[4] See also Hamlin and Wynn (2011) and Hamlin (2015) for further support of this idea.

[5] And so, it is also why the observed cycle is from discriminating cooperation to unconditional cooperation before cycling to defection.



d :	GG	GB	BG	BB
C	G	*	G	*
D	B	G	B	*
p :	C	D	C	**

Fig. 1 The leading eight reputation dynamics (*d*) and behavior strategies (*p*) from Ohtsuki and Iwasa (2004, 2006). At the top is a description of the scenario being judged, where "BG" stands for "someone with a bad reputation is the donor and someone with a good reputation is the receiver." In the BG interaction, it is good for the donor's reputation to cooperate (row 1) and bad to defect (row 2). The third row describes evolutionarily stable behavioral strategies for each possible pairing of good/bad donor and receiver. For example, the ESS in the BG interaction would be to cooperate. Concerning the asterisks, Ohtsuki and Iwasa say this: "Asterisk (*) is a wild card, implying that both G and B are included in the element… The [row] with double asterisks (**) is either C or D, which is determined according to the social norm that is adopted. It is C if and only if *d*(BB, C) = G and *d*(BB, D) = B; and it is D otherwise." Figure from Ohtsuki and Iwasa (2006), with permission from Elsevier. If the table were updated with the results of the Schmid et al. (2023) study (identifying the four most stable of the leading eight), the asterisk in the bottom right cell of the *d* table would be replaced with a B

reputation is public. However, Hilbe et al. (2018) showed that only one of the leading eight can actually evolve and sustain high levels of cooperation when these assumptions are relaxed (private reputation tracking with errors in assessment). Additionally, Ohtsuki and Iwasa only allowed categorical reputation scores, good versus bad. In a more recent paper, Schmid et al. (2023) worked with integer reputation scores (e.g., from −5 to 5) tracked privately by each simulated individual. These integer values are then translated into "good" or "bad" labels, which then guide the individual's actions in the game. With this wider possible range of character assessments, community agreement on reputation is easier to achieve, and a subset of the leading 8 strategies are much more likely to evolve, even against the "All defect" strategy (which is also an ESS under any set of parameters in Ohtsuki and Iwasa's simpler model of reputation assessments). Specifically, "norms must not be 'gullible': they should never assess a bad player defecting against another bad player as good (making them easily deceived by false shows of solidarity)." In other words, one should not assume that a bad person is defecting against another bad person for a good reason (i.e., that they "hate the bad" in Hurka's terms).

Critically, these modes of evaluating reputation roughly correspond with Hurka's notions of virtue and desert. The reputation dynamics correspond with different possible ways of determining virtue: assessing the character of an individual (as "good" or “bad”) based on their interactions with other agents (in accordance with the first two rows of Fig. 1). The behavioral strategies correspond with desert, or at least with how a person deserves to be treated (as opposed to what they deserve to receive, in general): if you are a “good” person, then other “good” agents “deserve” cooperation, and “bad” agents “deserve” defection (according to the *p* row in Fig. 1). Moreover, Hurka’s matching symmetry is consistent with this range of assessment rules and strategies. All of the leading eight combinations share a large portion of the intuitive matching rule: in all of the leading 8 dynamics, people are judged good if they cooperate with a good person and bad if they defect against a good person, and they are judged good if they are a good person defecting against a bad person.

However, it is easy to see that the resulting assessment and strategy space does not exactly reflect the most intuitive extension of Hurka’s recursive theory of virtue and desert. Intuitively, one might think Hurka’s theory would require a full matching rule (also called “stern judging”), which judges that it is always good to defect against bad actors and always bad to cooperate with them. However, Hurka’s theory is slightly more nuanced than this. Recall that in his theory, there are first-order goods in terms of which the second and third-order goods are recursively defined. There may be many first-order goods such as pleasure or knowledge, and in many cases, a good person might choose to cooperate with a bad actor in order to bring about those first-order goods. The amount of pleasure or knowledge that a bad person might receive from cooperation is a first-order good that may easily outweigh the third-order good of just deserts. Additionally, whether a bad person cooperates with or defects against a bad person, there is very little we can learn about whether they love or hate the bad. If all we know is that a bad person defects against a bad person, they may very well do so because they delight in the pain the bad person experiences, where pain has negative value at the first-order (even though a bad person experiencing pain can add a measure of positive value at the third-order level of desert).[6] Whereas, if we know the bad person cooperated with another bad person, they may very well be “loving” the third-order “bad” that exists when a bad person (second-order disvalue) experiences pleasure (first-order value). In other words, whether a bad person cooperates or defects with another bad person, we cannot easily conclude that they have turned over a new leaf and are now “loving the good.” So, the structure of Hurka’s recursive account of virtue and desert is consistent with the “non-gullible” norms of evaluation and action that are most likely to have evolved via indirect reciprocity.

Of course, empirical evidence is better support for claims about evolutionary products than any informal argument, and while such evidence is limited, it is suggestive. For instance, Swakman et al. (2016) put participants in anonymous iterated helping games and found that while individuals did vary in their particular strategies, their behavior was in line with the predictions of these models: a significant proportion of participants requested information about the reputation of receivers in the donation game before assessing the reputation of the sender, especially when the sender chose to defect. This is precisely what we would predict if the four most stable strategies differ according to whether cooperation is always good but agree that defection is bad unless it is against a bad person.

[6] Note how this parallels the discussion of “non-gullible” norms of evaluation above.

The point of the preceding discussion (of natural selection on these strategies) is to suggest that Hurka's intuitive symmetries between virtue and desert likely arose because of their effects (i.e., stabilizing cooperation) and advantages over other strategies (i.e., cognitive economy).[7] It is tempting to then say that desert and virtue evolved because of their instrumental value. For one, the evolutionary account seems to suggest that tracking reputation in this way is instrumentally rational for individuals. Moreover, cooperation itself is stabilized by the strategies we have discussed, and cooperation is usually quite beneficial to individuals. We might be tempted to think that the strategy arose because of its role in stabilizing cooperation, which benefits individuals. However, we have some doubts about these conclusions, and so, we think it a stretch to lump these selected effects under the heading of "instrumental value." On our evolutionary understanding of virtue and desert, these values arose from frequency-dependent selection, and as a result, their selection and maintenance does not derive from their instrumental value for individuals, at least not obviously so (even though the strategy does overlap with instrumental rationality in some contexts), nor did they evolve for their role in stabilizing cooperation. In models of indirect reciprocity, there are many stable equilibria (and many population states) that do not allow high levels of cooperation. Consider a monomorphic equilibrium composed entirely of unconditional defectors (which is also a stable equilibrium on these models). We would not say that this trait was selected for stabilizing defection. Such a claim would seem to assume that defection is a chosen end toward which evolution tends. Rather, we would say that the function of the dominant trait in any stable equilibrium (whether it stabilizes cooperation or not) is to prevent the spread of competing strategies. Nor would we assume that the trait of unconditional defection arose because of its instrumental value. In a population of discriminating cooperators, defection would certainly not be instrumentally valuable. Moreover, the benefits of unconditional defection are context dependent (beneficial in certain populations but not others), whereas the strategy itself is not. So, it is also plausible to suggest that the beneficiary of the trait is not actually the individual, since the individual playing this strategy would not benefit in many population states. Additionally, if the benefit of the strategy is best characterized as preventing the spread of competing variants, it would seem to be the *population* of unconditional defectors that most plausibly receives this benefit.

For these reasons, many who are familiar with these models would conclude that the leading eight strategy space also would not arise because of its role in stabilizing cooperation or for its instrumental value per se; that it makes more sense to say of these strategies that their function is to prevent the spread of competing strategies and that the beneficiary of the adaptation is actually the population of strategists, rather than the individual.[8] So, even though instrumental value is the usual contrast class for

[7] Cognitive economy would be significant because it could increase the basin of attraction for some of the leading eight over the others, at least under the assumption that a simpler rule to follow is more likely to get produced by random mutations than a more complex rule.

[8] Admittedly, it is possible to make the case that the leading eight would have evolved because of their instrumental value, even granting that (at equilibrium) unconditional defection would not have been selected in this way: Whereas unconditional defection does not vary its play across contexts, the leading eight do. For example, we can say that some of the leading eight strategies (at least the "non-gullible"

intrinsic value, we will henceforth say that the value of desert and virtue is likely to be *derived* instead of being intrinsic. To us, it seems clear that if our intuitions about desert and virtue are derived from their biological advantages (such as preventing the spread of competing strategies in a population), then we have a defeater for Hurka's claim that they are intrinsically valuable, or so we argue in the following section.

But first, there is one other wrinkle to iron out in the evolutionary story above. Hurka's work is based on evidence about our intuitions surrounding virtue and desert, whereas these evolutionary models are concerned with behavioral strategies. There is obviously a large theoretical gap between behavioral strategies and intuitions. However, we think it plausible that such a gap will eventually be filled. The patterns and dynamics uncovered by evolutionary game theoretic analyses are obviously at a high level of abstraction. If a recurrent social interaction realizes a game theoretic scenario that shapes generations of organisms to adopt a certain strategy, there must be a set of concrete, heritable implementations of that strategy. Moreover, it must be substantially flexible on the input and the output side. On the input side, this is because in our lineage, there is an extremely wide range of concrete scenarios that realize the structure of indirect reciprocity games. Any time someone chooses whether to give at a cost in a way that would provide a greater benefit to their partner, they are playing the indirect reciprocity game. Almost any form of cooperative aid fits this description: sharing tools, food, weapons, labor, even access to sex. On the input side, individuals must recognize each of these behaviors as cooperative or not, and on the output side, they need to be poised to update the reputation of the donor based on the prior reputation of the donor and receiver, share such information with others, and act in accordance to the correct rule when given the opportunity to provide aid when the donor in question becomes a potential receiver of their donation. This is clearly a high level of cognitive sophistication and a high level of automaticity. But intuitive processes just are processes that are highly automatic and cognitively sophisticated. In other words, it is difficult to imagine how these strategies could be realized through anything but an intuitive process.

ones that negatively evaluate defection of bad against bad) would also defect in a population of unconditional defectors, making such a strategy no less beneficial to the individual than the unconditional defection strategy across many population states. Thus, a case can be made that some of the leading eight would have evolved because of their instrumental value (read in terms of maximizing utility across contexts). We are not entirely convinced by this argument, because it is more parsimonious to give a univocal account of the function of traits that evolved by frequency-dependent selection: selection dynamics push populations to equilibrium states, and the frequency of traits that compose those states (whether they involve cooperation or defection or a polymorphic equilibrium of some kind) are "adaptive" (if that term even makes sense in this context) because they prevent substantial changes in the population state. Regardless of the contextual variability of the trait, the evolutionary dynamics that lead to its fixation in the population are the same in this important sense.

4 Debunking virtue and desert via selection-based explanations of intuitions

If all of these evolutionary considerations are accurate, then we evolved to treat virtue and desert as intrinsically good.[9] In other words, the most stable behavioral strategies in these models make character evaluations or actions independent of any perceived benefit. Regardless of other outcomes, a person cooperating with a good person is perceived as good, and giving a good person the cooperation they "deserve" is perceived as good (and so on). But the resulting intuitions or behavioral/evaluative tendencies evolved because of their outcome (e.g., preventing the spread of mutant strategies), as did the symmetry between virtue and desert. As such, these evolutionary explanations serve as *undercutting defeaters* for Hurka's intuitions about virtue and desert (Pollock, 1987; Klenk, 2018; Jaquet, 2022).

To reiterate, Hurka's thought (summarized above in Sect. 2) seems to be that virtue and desert share structural properties and that this structural similarity provides some evidence that they are good in themselves, independently of first-order goods. Moreover, moral intuitions suggest that virtue and desert are good in themselves. Nevertheless, these intuitions, and the symmetry between them, have derivative value and evolved because of their derivative value. Thus, they cannot be evidence that virtue and desert have non-derivative (i.e., intrinsic) value. It would only be a coincidence if the symmetry and our intuitions happen to coexist with the intrinsic value of virtue and desert.[10]

Consider an analogy: Let's say that you hold the intuition that your friend Monica is a robot.[11] However, one day, you receive reliable information that you were given a pill that was designed to make you believe that Monica is a robot. As a result of learning this information, you should not necessarily conclude that your intuition is false, for it could still be true that Monica is a robot (perhaps by coincidence). However, you should nevertheless accept that the intuition is not good evidence for the belief, for you have learned that the process that generated the intuition was unreliable.

In this example, the information about the pill serves as an undercutting defeater (Pollock, 1987; Klenk, 2018; Jaquet, 2022); that is, a reason why the intuition is the result of an off-track process. Our evolutionary debunking argument works in a similar fashion. It introduces a causal claim: some set of intuitions were significantly shaped by natural selection. Next, it puts forward an epistemic claim: this causal relation undercuts the evidential value of the intuition. The result: the intuitions are not good evidence.[12]

[9] This is obviously not to say that we evolved to have de dicto beliefs (e.g., that virtue is intrinsically good) at this level of abstraction. However, our intuitive responses to the virtues we see in others may lead to the development of de dicto intuitions at lower levelis of abstraction (e.g., that courage is good in itself). One way in which these intuitions could solidify is through a process of rationalization (e.g. Cushman, 2020).

[10] See Wiegman (2017: Sect. 3.2 and 3.3) for a more detailed, analogous argument concerning the intrinsic value of retribution.

[11] This thought experiment was inspired by Joyce (2006).

[12] For examples of this kind of argument being used to debunk certain sets of our moral intuitions and beliefs, see Singer (2005), Street (2006, 2016), Joyce (2006), Greene (2014), de Lazari-Radek and Singer (2012), Vavova (2015), Wiegman (2017), Sauer (2018), and Jaquet (2022).

In response to this debunking argument, some might worry about the risk of overgeneralization (Kahane, 2011; Berker, 2014). That is, while we may be arguing that a certain limited set of intuitions concerning virtue and desert can be explained by natural selection, who's to say that evolutionary explanations stop there? Indeed, Street (2006) and Joyce (2006) claim that *all* of our moral beliefs and intuitions can likely be explained by natural selection, and thus we have reason to accept more general skepticism.

While these kinds of *global* debunking arguments are certainly interesting and plausible, it is important to understand that most philosophers working on evolutionary debunking arguments admit that they are not relying on specific empirical evidence. Instead, they *assume* that certain intuitions were shaped by natural selection, and then go on to discuss the philosophical implications. For instance, Street (2006, Sect. 3) writes, "[I]t must suffice to emphasize the hypothetical nature of my arguments, and to say that while I am skeptical of the *details* of the evolutionary picture I offer, I think its *outlines* are certain enough to make it well worth exploring the philosophical implications." More pointedly, Vavova (2014: 79) asserts that, "No one, not even the debunker, thinks [the empirical claim] is conclusive."

In this paper, we are taking a different approach. We do not rely on vague assumptions about the evolution of morality. Instead, we put forward a very specific empirical case concerning a certain set of intuitions. Due to this, it is not necessary to worry about the risk of overgeneralization. We only show that intuitions about the intrinsic value of character and desert are defeated. Our evidence points to these specific intuitions being selected for their biological benefits, and the fact that their value is derived from their selected effects makes the process unreliable as an indicator of intrinsic value.[13] Of course, it may be true that a broader set of our intuitions and beliefs can be explained by natural selection (and our account would not be opposed to such a broader explanation), but as of now such an empirical case has not been made, and thus, the risk of overgeneralization should not be seen as a worry for our account.[14]

[13] As Jaquet (2018) points out in response to Kahane (2014), there is another route open to a would-be vindicator of intrinsic value: the subjectivist route. On a subjectivist view of intrinsic value, something like pleasure or virtue would be intrinsically valuable because the subject desires it (or has some other pro-attitude toward it) for its own sake. The evolutionary story is then irrelevant: no explanation of *how* the subject came to desire virtue for its own sake undermines the fact that she does desire it for its own sake. So, on a subjectivist account, virtue and desert still hold intrinsic value even if our evolutionary account of virtue and desert is correct. While important to consider, this point does not help Hurka's account, for Hurka appears to be offering an objective account of intrinsic value. (Although the recursive structure of it, sheared of Hurka's inferences about objectivity, can be adapted to subjectivist theories of value such as Nietzsche's, see, e.g., Sinhababu, 2015, p. 288 – 89). As described in Sects. 1 and 2 of this paper, Hurka appears to accept that there are certain objective goods, and goes on to add other objective goods via recursion (1992: 150–152; 2001: 8–9). Thus, our debunking argument is primarily directed at his and other *objectivist* accounts of the intrinsic value of virtue and desert, and we doubt that it can be extended to subjectivist accounts. Thanks to an anonymous reviewer for raising this concern.

[14] Jaquet (2022) makes a similar point about the threat of overgeneralization.

5 Conclusion

Thomas Hurka's recursive account of value appeals to certain intuitions and symmetries to expand the class of intrinsic values. In this paper, we argued that these can be explained by theories of indirect reciprocity in evolutionary biology and evolutionary game theory. Indeed, our tendencies to value virtue and desert as intrinsic goods are better explained by their value for solving certain selection problems. This shifts the burden of proof in favor of the hypothesis that virtue and desert have only derivative value.

We do not intend this as a wholesale criticism of Hurka's work on the symmetries between virtue and desert. While we have argued against some of the conclusions he draws from them, his observations and descriptions of these symmetries are significant contributions to our understanding of value. From trolley dilemmas to moral luck cases to intuitions about desert and virtue, it seems to us that philosophers often have a critical role to play in empirical moral psychology: identifying patterns in moral beliefs and action that cry out for explanation. In some cases, the phenomenon in question will be best explained by evolutionary considerations. In others, the explanation may be cultural or psychological, or perhaps even structural or normative. Our suspicion is that any time an intuition seems obviously true but difficult to explain on the basis of experience (as is the case with the intuitions we consider here), an evolutionary explanation is a good candidate to consider.

Acknowledgements Thanks to two anonymous referees for Synthese. Their thoughtful and constructive feedback helped us clarify and improve our arguments.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aristotle (1984). Nicomachean Ethics. In *The Complete Works of Aristotle*. Trans. Ross, W. D. and Urmson, J. O., ed. Barnes, J. Princeton University Press.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Berker, S. (2014). Does evolutionary psychology show that Normativity is Mind-Dependent? In J. D'Arms, & D. Jacobson (Eds.), *Moral Psychology and Human Agency Philosophical Essays on the Science of essays* (pp. 215–252). Oxford University Press.

Boyd, R., & Lorderbaum, J. (1987). No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature*, *327*, 58–59.

Bshary, R., & D'Souza, A. (2005). Cooperation in communication networks: Indirect reciprocity in interactions between cleaner fish and client reef fish. In P. McGregor (Ed.), *Communication networks* (pp. 521–539). Cambridge University Press.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, e28.

de Lazari-Radek, K., & Singer, P. (2012). The objectivity of Ethics and the Unity of Practical Reason. *Ethics*, *123*(1), 9–31.

Greene, J. (2014). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics*, *124*, 695–726.

Greiner, B., & Levati, M. (2005). Indirect reciprocity in cyclical networks: An experimental study. *Journal of Economic Psychology*, *26*(5), 711–731.

Hamilton, W. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, *7*(1), 17–52.

Hamlin, J. (2015). The case for social evaluation in preverbal infants. *Frontiers in Psychology*, *5*, 1563.

Hamlin, J., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*, 30–39.

Hilbe, C., Schmid, L., Tkadlec, J., Chatterjee, K., & Nowak, M. A. (2018). Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the National Academy of Sciences*, *115*(48), 12241–12246.

Hu, Y., Ma, J., Luan, Z., Dubas, J., & Xi, J. (2019). Adolescent indirect reciprocity: Evidence from incentivized economic paradigms. *Journal of Adolescence*, *74*, 221–228.

Hurka, T. (1992). Virtue as loving the good. *Social Philosophy and Policy*, *9*(2), 149–168.

Hurka, T. (2001). The common structure of virtue and desert. *Ethics*, *112*(1), 6–31.

Jaquet, F. (2018). Evolution and utilitarianism. *Ethical Theory and Moral Practice*, *21*, 1151–1161.

Jaquet, F. (2022). Speciesism and tribalism: Embarrassing origins. *Philosophical Studies*, *179*, 933–954.

Joyce, R. (2006). *The evolution of morality*. MIT Press.

Kahane, G. (2011). *Evolutionary Debunking Arguments Nous, 45*, 103–125.

Kahane, G. (2014). Evolution and impartiality. *Ethics*, *124*(2), 327–341.

Kant, I. (1899). *Critique of pure reason*. Colonial.

Kato-Shimizu, M., Onishi, K., Kanazawa, T., & Hinobayashi, T. (2013). Preschool Children's behavioral tendency toward Social Indirect Reciprocity. *Plos One*, *8*(8), e70915.

Klenk, M. (2018). Objectivist conditions for defeat and evolutionary debunking arguments. *Ratio*, *32*(4), 246–259.

Meristo, M., & Surian, L. (2013). Do infants detect indirect reciprocity? *Cognition*, *129*(1), 102–113.

Moore, G. E. (1993). *Principia ethica*. Cambridge University Press. https://philpapers.org/rec/MOOPEE

Nowak, M., & Sigmund, K. (1998a). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, *194*, 561–574.

Nowak, M., & Sigmund, K. (1998b). Evolution of indirect reciprocity by image scoring. *Nature*, *393*, 573–577.

Nowak, M., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, *437*, 1291–1298.

Ohtsuki, H., & Iwasa, Y. (2004). How should we define goodness?—reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology*, *231*, 107–120.

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*, 435–444.

Pollock, J. (1987). Defeasible reasoning. *Cognitive Science*, *11*(4), 481–518.

Romano, A., Balliet, D., & Liu, J. (2017). Parochial trust and cooperation across 17 societies. *Proceedings of the National Academy of Sciences*, *114*(48), 12702–12707.

Ross, D. (2003). *The right and the good*. Oxford University Press.

Sauer, H. (2018). *Debunking arguments in ethics*. Cambridge University Press.

Schmid, L., Ekbatani, F., Hilbe, C., & Chatterjee, K. (2023). Quantitative assessment can stabilize indirect reciprocity under imperfect information. *Nature Communications*, *14*(1), 2086.

Seinen, I., & Schram, A. (2006). Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review*, *50*(3), 581–602.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*, 331–352.

Sinhababu, N. (2015). Zarathustra's metaethics. *Canadian Journal of Philosophy*, *45*(3), 278–299.

Street, S. (2006). A darwinian dilemma for realist theories of value. *Philosophical Studies*, *127*, 109–166.

Street, S. (2016). Objectivity and Truth: You'd Better Rethink It. In R. *Shafer* Landau (Ed.) *Oxford Studies in Metaethics, vol. 11*. Oxford University Press.

Swakman, V., Molleman, L., Ule, A., & Egas, M. (2016). Reputation-based cooperation: Empirical evidence for behavioral strategies. *Evolution and Human Behavior*, *37*, 230–235.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 35–57.

van Apeldoorn, J., & Schram, A. (2016). Indirect reciprocity; a field experiment. *PLOS ONE*, *11*(4), 1–11.

Vavova, K. (2014). Debunking Evolutionary Debunking. In R. Shafer, & Landau (Eds.), *Oxford Studies in Metaethics, vol. 9*. Oxford University Press.

Vavova, K. (2015). Evolutionary debunking of Moral Realism. *Philosophy Compass*, *10*(2), 104–116.

Wedekind, C., & Milinski, M. (2000). Cooperation through image scoring in humans. *Science*, *288*(5467), 850–852.

Wiegman, I. (2017). The evolution of retribution: Intuitions undermined. *Pacific Philosophical Quarterly*, *98*(2), 193–218.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.