ISSN 1897-1652

POLISH JOURNAL OF PHILOSOPHY

Volume VIII, No. 2 (2014)

Jagiellonian University
Kraków

POLISH JOURNAL OF PHILOSOPHY
FALL 2014

Articles

Discussions

Book Reviews

POLISH JOURNAL OF PHILOSOPHY
Vol. VIII, No. 2 (Fall 2014), 65-74.

Harming Yourself and Others: a Note on the Asymmetry of Agency in Action Evaluations

Erich Rast
IFILNOVA Institute of Philosophy
Universidade Nova de Lisboa, FCSH

Abstract. Principles are investigated that allow one to establish a preference ordering between possible actions based on the question of whether the acting agent himself or other agents will benefit or be harmed by the consequences of an action. It is shown that a combination of utility maximization, an altruist principle, and weak negative utilitarianism yields an ordering that seems to be intuitively appealing, although it does not necessarily reflect common everyday evaluations of actions.

1. Introduction

It is a fact of everyday life that we make a difference in our moral assessments between the case when an acting agent harms or benefits himself and the case when other people are harmed or benefit from the action. The goal of this note is to identify a small set of general criteria that may account for this asymmetry and discuss their relevance to broadly-conceived utilitarian decision-making. Since the account is minimalist, no claim is made that these principles can or should be used in actual decision-making. There are so many more aspects of agency, for example the action/omission distinction (Baron, 1996), the notion of moral hazard and the principal-agent problem (Spence, Zeckhauser, 1971) in economics that lead to the development of agency theory (see Eisenhardt, 1989, Shapiro 2005 for overviews), or the Knobe effect (Knobe, 2003ab), that many more ingredients would be needed apart from the ones discussed in this short note in order to obtain a reasonable understanding of the effects of agency in a broadly-conceived utilitarian setting.

One of the assumptions of this article is that there is indeed a broad, though perhaps not universal, consensus that there is an asymmetry insofar as agency is concerned, so some examples should be given first. Knowingly harming yourself has a different moral status than if you knowingly harm others, and the difference may be thought to be particularly huge if one and

DOI: 10.5840/pjphil20082219

the same action results in harm or benefits to you and others at the *same* time. Somebody who gains a lot by his actions while harming others is considered wicked. Conversely, exposing yourself to unavoidable harm when someone else benefits from your action is often considered commendable. For example, a 'good soldier' is supposed to endure a lot of suffering for the greater good of the people he is fighting for. As another example, a firefighter who dies while attempting to save others will be praised as a hero, although he is generally expected to guard his life as best as he can. In contrast to this, a politician who orders someone else to take a huge risk in order to further his own agenda will likely not be held in such high regard once things have gone awry. The question is whether it is possible to find general principles to guide such assessments and how compatible these are with common everyday assessments.

2. Maximizing Utility and the Weak Negative Utilitarian Principle

Suppose that an action can do harm to yourself or others (H), you or others may benefit from it (G), or the action is neutral (N). This discrete conceptualization should be understood as an abstraction from the underlying threshold-based utility assessments. Harm and benefits may come in relative or absolute variants. Relative harm is done when the utility after an action minus the utility before an action exceeds a negative threshold. Relative benefits are obtained when the utility after an action minus the utility before an action exceeds a positive threshold. Finally, an action is neutral when neither of these cases obtain. A notion of absolute harm can be formulated by looking at the difference between the utility after an action and a negative threshold. If the difference is negative, the corresponding amount of harm has been done, provided that the utility before the action was above the negative threshold (otherwise the harm had been done earlier). The definition of absolute benefits is analogous for a positive threshold, and the neutral case is defined as before. All of this can be formulated in a precise way in a multi-agent utilitarian framework and it can be combined with the standard treatment of risk and Expected Utility theory, but these details do not matter for the following discussion, which, for the sake of simplicity, only takes into account abstract orderings based on the H/N/G distinction. Naturally, this discrete distinction only makes sense as long as the respective levels of harm and benefits are on a par. It is therefore from now on assumed that any amount of harm under discussion is roughly within the same lexical comparison class as the corresponding level of benefits, since otherwise a full quantitative comparison would be needed. So, for instance, if G stands for a life saved, then H cannot stand

for a pinprick. Conversely, if H stands for a possibly avoidable death, then G cannot stand for the fact that somebody gets a chocolate bar. However, deaths and lives are commensurable, and pinpricks and chocolate bars are also commensurable. (While this is a controversial assumption, some weighing of harm against benefit will always be necessary, or else paradoxes might occur and many principles such as the 'double effect principle' could no longer be formulated at all.) It would also be possible to use multiple thresholds and thereby distinguish strong from weak harm and benefits, but to keep things simple only the H/N/G distinction will be used in what follows. Suppose, as a further simplification, that only one other agent is affected by the acting agent's action. Again, this is not a principal limitation, but only simplifies matters.

Under these assumptions – that the H/N/G distinction is taken as a basis, and that only the acting agent and another agent who is affected by the action are considered – an action can have nine possible effects: GG, GN, NG, NN, NH, HN, HH, HG, GH. Let the first letter stand for the acting agent and the second one for the other agent. Then, for instance, NG means that someone's action is neutral for himself and beneficial to someone else. To give another example, HG means the acting agent harms himself and in doing so does good to someone else. One may now ask how these nine options should be ordered, and according to what criteria such an ordering should take place. Ideally, these criteria ought to be minimal and give some insights into common intuitive assessments of actions. If they can be found for the H/N/G distinction, it seems likely that an analogous answer could be given for a more complicated account based on continuous utility functions with thresholds.

In a broadly-conceived utilitarian setting, utility maximization must somehow enter the picture. The combination GG should always be on top, because it represents the optimal case when someone obtains benefits for himself and for others by the same action, which is truly a win-win situation. Or so one might think. By the same token, HH should always be the least preferred course of action, for you cannot do any worse than harming yourself and others at the same time. These considerations suggest the following principles:

(OPT) Win-Win Principle:

An action in which all involved parties win more than the benefit threshold ('win-win') is preferable to all other outcomes.

(PES) Lose-Lose Principle:

An action that harms everyone ('lose-lose') is the least preferable.

If one buys into utility maximization as a general principle in the first place, then the following variant will subsume the two previous principles:

(MAX) Utility Maximization Principle:

Aim to maximize overall utility.

This principle has been deliberately left vague in order to illustrate that utility maximization can be understood in slightly different ways. As one way to make it more precise in the current simplified and discrete setting, scores of equal weight may be attributed to the H/N/G classification, regardless of the question of who acted for whom. Suppose G=1, N=0, and H=-1. This yields the following table:

	G	N	H
G	2	1	0
N	1	0	-1
H	0	-1	-2

From this table, a weak ordering can be read off:

(1) GG > NG ~ GN > HG ~ NN ~ GH > NH ~ HN > HH

Let us call this ordering the 'unweighted sum account', because it puts equal weight on benefits and harm. In a quantitative setting, this approach corresponds to an additive model in which utilities are summed up and possibly divided by the number of utilities to normalize the measure. Principles OPT and PES are faithfully represented in this account, but in other respects the unweighted sum account is hard to accept. Option *NN* is clearly preferable to GH or HG, because an NN action is neutral; it might result in a little bit of loss or gain to some of the agents involved, but does not do any harm, whereas the other two actions cause harm to someone. It seems that harm should be avoided more strongly. Setting G=1, N=0 and H=-2 we obtain the following table:

	G	N	H
G	2	1	-1
N	1	0	-2
H	-1	-2	-4

From this table a slightly different weak ordering is obtained:

(2) GG > GN ~ NG > NN > HG ~ GH > HN ~ NH > HH

This time, *NN* is preferred to *HG* and *GH* which matches more closely our common judgments about harm. If one considers this ordering more acceptable than the previous one, the following principle should be endorsed:

(WNU) Weak Negative Utilitarian Principle:

Avoiding harm is more desirable than obtaining benefits.

This should be interpreted only in the way indicated above, that harming someone counts more than obtaining a benefit, because a stronger reading may lead to the paradoxes of Negative Utilitarianism laid out by Smart (1968) and Ord (2013), among others. Notice that this principle cannot count as an ordering principle that can stand on its own, because taken by itself it would lead to the following, rather uninformative ordering:

(3) GG ~ GN ~ NG ~ NN > HG ~ GH ~ HN ~ NH ~ HH

However, WNU can be combined with MAX by first applying the latter, and then disambiguating with the former whenever possible, resulting in (2). So we have a meta-order and this combination of the two principles provides a reasonable weak ordering, which is, technically speaking, a preorder relation. Given (2), one might ask how a linear ordering can be obtained by applying further principles. As it turns out, answering this question at the proposed level of generality is more complicated than it might seem at first glance. This is the topic of the next section

3. Egoism, Altruism, and the Rule Hierarchy

One way to obtain a linear ordering is by combining the following flawed principle with utility maximization:

(EGO) 'Me First' Principle:

Prioritize the self-benefits of the agent over the benefits of others, and, correspondingly, avoid harm to yourself before avoiding harm to others.

As it stands, the principle appears to be flawed, and indeed it leads to undesirable results when taking precedence over joint utility maximization, i.e., when the agent seeks to always maximize his own gains and only after this principle has been applied agent-agnostic maximization is used to sort out the remaining cases. EGO alone boils down to

(4) GG ~ GN ~GH > NG ~ NN ~ NH > HG ~ HN ~ HH

So the resulting combined linear ordering is:

(5) GG > GN > GH > NG > NN > NH > HG > HN > HH

Clearly, the combination EGO > MAX is the hallmark of evil egoism and therefore wholly unacceptable. A person with preferences like GH > NG or GH > NN is rightly considered wicked, someone who is willing to sell his own grandmother. What about a more decent form of egoism? Suppose MAX > WNU is given precedence over the 'Me First' Principle instead, meaning that EGO only serves as a tie-breaker for (2). The resulting linear ordering is then:

(6) GG > GN > NG > NN > GH > HG > NH >HN > HH

While this looks better than (5), GH > HG might still be an offender. Is it? The problem is that if we consider an unavoidable choice between two acts, one with result GH and one with a result in {HG, NH, HN, HH}, then the egoist's preference might seem alright. If GH is considered in isolation, though, then the action seems to be wicked. Many laws actually disallow GH-type actions provided that the harm involved is high enough, no matter how large the corresponding benefit is to the delinquent.

Even if one takes (6) to be acceptable, there is another alternative to consider. Parallel to EGO, an altruist principle may be introduced:

(ALT) 'Others First' Principle:

Prioritize the benefit to others over self-benefits, and, correspondingly, avoid harm to others first before avoiding harm to yourself.

This principle corresponds to the weak ordering:

(7) GG ~ NG ~ HG > GN ~ NN ~ HN > GH ~ NH ~ HH

While the second part of the principle sounds strange, since always preferring to harm yourself over harming others does not seem to improve a person's survival capabilities, it is noteworthy that the linear ordering resulting from prioritizing ALT > MAX does not look nearly as bad as (5):

(8) GG > NG > HG > GN > NN > HN > GH > NH > HH

Problematic in this is HG > GN , though. The alternative HG is a huge sacrifice, and the recommendation to let go of GN in favor of HG might be hard to swallow for anyone but a die-hard altruist. Perhaps GN and NG are the sane variants of egoism and altruism, whereas GH and HG are the insane versions. (Remember that we are talking about reasonably high levels of harm and benefits.) Be that as it may, a better result can be obtained by using ALT only as a tie-breaker for (2), resulting in:

(9) GG > NG > GN > NN > HG > GH > HN > NH > HH

This is the winning candidate for an ordering of possible acts in this abstract fashion. It is based on three simple principles with rule priorities MAX > WNU > ALT and should be intuitively acceptable to anyone but the wicked and social Darwinists.

There is an alternative to this principle based on Mill (1859, 1869) who wrote that "... the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others." (Mill 1869, p. 9) While this famous passage describes a deontic restriction of liberty, it can be modified into the following utilitarian rule:

(MIL) No Harm Principle:

Do not harm others if at all avoidable.

To this principle corresponds the weak ordering:

(10) GG ~ GN ~ NG ~ NN ~ HG ~ HN > GH ~ NH ~ HH

Since it leaves HG and NN in one partition, WNU is needed this time. However, the rule hierarchy MIL > MAX > WNU still does not lead to a linear ordering:

(11) GG > GN ~ NG > NN > HG > GH > HN > NH > HH

The combination GN ~ NG cannot be resolved by MIL, and the same holds for the rule hierarchy MAX > WNU > MIL. But perhaps this is exactly how things ought to be. As long as nobody is harmed, acting for your own benefit should be treated as being on a par with acting for the benefit of others.

Considering that they are based on such simple principles, it is puzzling that neither of the above orderings necessarily reflects our common everyday judgments. It seems that we often violate even the Win-Win Principle in everyday assessments of actions, both prior to acting and in after-action evaluations. For example, a nun who gives up her own good for the sake of helping starving children, leading to a life of many HG-decisions, is sometimes praised more than a talented and wealthy businessman who manages to achieve many GG actions in his continuous striving for Pareto optimality, and it is easy to come up with similar examples for HG > GN. These preferences directly violate MAX and also violate harm avoidance principles like WNU and its stronger variants. There is perhaps a virtue-ethical reason why we tend to value HG-actions higher than GN- or GG-actions: The former leave no doubts about the proper motives of the agent, whereas the latter allow for a purely egoistic interpretation. When someone runs into a burning house to save a child and manages to bring it to safety only to later succumb to his own injuries, a clear-cut HG action, we consider him a hero. If, on the other hand, a board member of a big pharmaceutical company reaches out to help a dying child whose parents cannot afford the expensive medication, and by this action gets so much publicity that the company's stock options rise and millions can be saved on advertising, opinions about the action may be divided. Helping the child is good, but doubts about the motives remain.

4. Concluding Remarks

Some principles have been identified that in the appropriate order of application lead to an intuitively acceptable ordering of action alternatives categorized by the H/N/G scheme. One of them that might appear to reflect a utilitarian spirit, principle EGO, had to be rejected, because it leads to highly counterintuitive orderings. Moreover, none of the proposed principles clearly mirrored our everyday assessments. While it seems that the H/N/G distinction is too coarse-grained for an adequate evaluation of actions in general, and it has not been claimed that it could serve this purpose, it seems worthwhile to take a closer look at the ordering principles that might underlie our everyday assessments when additional factors are

taken into account. Is it perhaps possible to find rational criteria that result in HG > GG? If so, then such an ordering cannot be based on utility maximization or a combination of the other principles discussed in this article, and so my overall conclusion is negative: If there is a rational justification in a broadly-conceived utilitarian setting for the above assessments of actions that cannot be explained by the H/N/G scheme, then it must hinge on other, hitherto unknown factors.

Since the H/N/G distinction is limited to comparisons between levels of benefits and harm that are clearly on a par, future research must address the question of how similar principles can be formulated directly in terms of utility functions that allow for distinguishing the acting agent from other persons affected by the action. These must be based on negative and positive thresholds for levels of harm and benefits respectively, in order to account for the neutral case that represents very small gains and losses. The principles MAX, EGO and ALT seem to be easily expressible in such a framework, but WNU and a combination of all of them might pose problems.

Acknowledgements
I am very grateful for the grant SFRH/BPD/84612/2012 of the *Portuguese Foundation for Science and Technology* under which the work on this article was conducted.

References

Baron, J. (1996). Do No Harm. In Messick, D. M., Tenbrunsel, A. E. (eds.) *Codes of Conduct: Behavioral Research into Business Ethics* (pp. 197-213). New York: Sage Foundation.

Eisenhardt, K. (1989). Agency Theory: An Assessment and Review. *Academy of Management Review 14 (1)*, 57-74.

Knobe, J. (2003a). Intentional Action and Side Effects in Ordinary Language. *Analysis 63*, 190-193.

Knobe, J. (2003b). Intentional Action in Folk Psychology: An Experimental Investigation. *Philosophical Psychology 16*, 303-324.

Mill, J. S. (1869). *On Liberty*. London: Roberts & Green, London (first publ. 1859).

Ord, T. (2013). Why I'm Not a Negative Utilitarian. University of Oxford, published online at URL http://www.amirrorclear.net/academic/ideas/negative-utilitarianism/index.html. Date of retrieval: 20. 2. 2014.

Shapiro, S. (2005). Agency Theory. *Annual Review of Sociology 31*, 263-284.

Smart, R. N. (1958). Negative Utilitarianism. *Mind 67*, 542-543.
Spence, M. & Zeckhauser, R. (1971). Insurance, Information, and Individual Action. *American Economic Review 61*, 380-387.