

On the Role of Changing Evaluations in Normative Theories

—Summary—

Sebastian Lutz*

2009–01–13

Abstract

I argue that the formalism of rational decision theory can accommodate many kinds of evaluations, whether determined by preferences, happiness, or successful application of abilities. I generalize the resulting abstract theory to allow for an agent's evaluation to change as a result of her actions. Because of its abstract nature, the generalization captures the point of agreement between previous treatments of changing evaluations by Phillip Bricker and Krister Bykvist. The generalization allows a rational criticism of evaluations and, since it can be extended to include moral theories, a moral criticism of evaluations as well. When the evaluation is determined by an agent's preferences, the generalization allows solutions to the problems of internalized oppression, negative utility hogs, and the alienation objection to impartial moral theories.

1 Introduction

Many teleological normative theories, rational, prudential, and ethical, depend for the evaluation of outcomes of actions—and hence for the inference of their normative statements—on features of the world. Evaluations of outcomes can, for example, be determined by preferences, rational preferences, happiness, or achievements. These determinants can change, and are sometimes even changed by the very actions that the theories demand.

The dependence of a normative theory on evaluations leads to at least two problems: The first is that, if the evaluations change, it is not clear which of the evaluations is the right one, because there can be different evaluations at different times and even different evaluations in each of the possible outcomes of an action. The second problem is independent of the evaluations' change:

*Theoretical Philosophy Unit, Utrecht University and Department of Philosophy, University of Western Ontario. Parts of this paper were written during a fellowship at the Center for Logic and Philosophy of Science, Tilburg University.

There does not seem to be a way to argue that someone's evaluation is imprudent, irrational, or unethical based on the normative theory that depends on that very evaluation.

Some suggestions have been made to take changing evaluations into account, notably by Bricker (1980) and Bykvist (2006), and I want to show in my talk that there is an abstract theory of changing evaluations that accommodates their results and gives sensible generalizations of the above laundry list of normative theories. Conveniently, this abstract theory then solves the two mentioned problems and also, for some kinds of evaluations, the seemingly unconnected problem that impartial ethical theories are alienating because they demand an intellectual calculation in situations where a compassionate reaction is intuitively appropriate.

2 An abstract theory for changing utility functions

Decision theory typically assumes a utility function u_k from states to (tuples of) real numbers, determined by preferences of an agent k . Decision theory's normative claim then lies in the demand to maximize the expected utility

$$e_d(a) = \sum_s P(s|a) u_k(s), \quad (1)$$

a sum over the probabilities of all possible outcomes s of an action a , weighted with their conditional probability given a . The assumptions that underly this equation are weak enough to allow a variety of other determinants, for example happiness, the success of the application of abilities. Each of these kinds of determinants leads to a different kind of evaluation, but their formal features are often the same.

For example, while Bricker (1980) and Bykvist (2006) disagree on the determinants and consequently on the kind of evaluation, I will show that they agree on the treatment of changing evaluations: The value of the outcome of an action is determined by the evaluations that hold for that outcome, not some other possible outcome of the action. Bricker and Bykvist assume deterministic outcomes of actions, and while Bricker makes the strong assumption that the value of any situation in one's life is determined by one's evaluations at all times, Bykvist considers only the special case where there is only one change of the evaluation during one's life. I generalize their results to probabilistic changes of the evaluation and any set τ of times t at which the evaluation is considered relevant. This leads to the formula

$$e(a) = \sum_{t \in \tau} \sum_{u_{k,t}} \sum_s P(u_{k,t} \wedge s | a) u_{k,t}(s) \quad (2)$$

for the expected value of an action a . ' $u_{k,t}$ ' in the conditional probability is to be interpreted as the statement that for agent k , the evaluation at time t is $u_{k,t}$. This generalization of the standard equation (1) takes into account that different actions can bring about different utility functions and that utility functions at times other than the time of action may be relevant for the value of s .

I argue that for a variety of determinants other than preferences, the value of a state s at time t_s is determined by the evaluation u_{k,t_s} . Even when the determinants are preferences, u_{k,t_s} plays

a role except for the special case that k 's preferences are constant over time, only preferences at the time of action t_a are relevant, and at t_a , k does not have a preference for future preference fulfillment. More general preferences about preferences can be taken into account, but do not change the results of my analysis.

In combination with the demand to maximize the expected value of an action, the generalized formula (2) for expected values recovers intuitively plausible demands, like the one to avoid irrational fears and preferences that cannot be fulfilled. These results do not fundamentally change for other sets τ of times as long as $t_s \in \tau$.

3 Ethical theories

Ethical theories that restrict the possible actions of k without using the evaluation or its determinants can easily be incorporated into the analysis by demanding the maximization of the expected value (2) of an action allowed by the theory. When the evaluations are determined by preferences, this already leads to a response to the alienation objection, according to which impartial moral theories do not allow for empathic motivations of morally right actions: Prudence demands that one prefers those actions that one has to do anyway, so one is always motivated to do the right thing.

Since there are many ways to incorporate the evaluation into an ethical theory, I will just treat utilitarianism as one example. The typical step from decision theory to utilitarianism consist in summing over all agents. The same step leads from the generalized formula (2) to

$$e_U(a) = \sum_k \sum_{t \in \tau} \sum_{u_t} \sum_{s \in S} P(u_{k,t} \wedge s | a) u_{k,t}(s). \quad (3)$$

This generalized formula for utilitarianism already solves the problem of the negative utility hog (the realistic version of Nozick's utility monster), an agent k whose utility function is so excessively negative for some states s that everyone has to make sure s never comes about, even though, intuitively, k 's dislike of s is just a whim. Demanding from k to maximize the generalized expected utility (3) means demanding k to change her preferences if it is comparably easy for her, that is, if her dislike of s is indeed just a whim.

The response to the alienation objection holds here as well, and maybe even with more force, because the change of preference is even morally demanded.

4 Final remarks

The generalized formulas presented here hold not only for preference based normative theories, but for any theory relying on an ordering in the same way decision theory does. Because of this, the results of the analysis are rather robust. A change of the interpretation of the evaluations does not invalidate the results. The analyses by Bricker and Bykvist are an example: Although based

on different determinants of the evaluation, both their positions are captured by the abstract theory here.

References

Phillip Bricker. Prudence. *The Journal of Philosophy*, 77(7):381–401, July 1980.

Krister Bykvist. Prudence for changing selves. *Utilitas*, 18(3):264–283, 2006.