Hard-to-reverse decisions destroy option value

Stefan Schubert

stefan.schubert@centreforeffectivealtruism.org

Ben Garfinkel

ben.garfinkel@centreforeffectivealtruism.org

Centre for Effective Altruism

Summary: *Some strategic decisions available to the effective altruism movement may be difficult to reverse. One example is making the movement's brand explicitly political. Another is growing large. Under high uncertainty, there is often reason to avoid or delay such hard-to-reverse decisions.*

Introduction

The importance of option value is widely appreciated within the effective altruism movement. In an uncertain world, keeping multiple paths open can be very valuable.

One aspect of option value is *reversibility*. When we consider a change from the *status quo* to a new state of affairs, we risk losing option value if the decision is difficult to reverse. If we have a white cloth, we can dye it black at any time. However, once we have dyed it black, it would take much more work to make it white again.

The effective altruism movement faces many similar strategic situations, where it is easier to leave the *status quo* than to get back. For instance:[1]

- The effective altruism movement is currently not strongly associated with any political party, or with any of the mainstream political ideologies. In that sense, the

[1] Note that the primary point of these examples is to illustrate the notion of reversibility. There might be reasonable disagreement on how reversible these strategies are.

effective altruism brand is relatively apolitical. This could easily change, e.g., if the movement allied itself with a certain political party. However, once that step has been taken, it might be hard to go back to an apolitical brand.

- The effective altruism movement is currently quite small. The movement may try to grow big, but if it does, it will arguably be hard to reverse that decision. Shrinking the movement in a way which does not cause serious damage is presumably very difficult.

- The effective altruism movement has invested in acquiring a reputation for integrity, rigour, friendliness, and other kinds of prosocial behaviour. It may, however, decide that such a reputation is too costly to uphold, and that some level of dishonesty, lack of rigour, or unfriendliness is acceptable. Since it is easier to destroy a good reputation than to build one, it is plausibly hard to reverse such a decision.[2]

What is reversibility?

To understand the notion of reversibility, let us first look at the notion of option value. It can be defined as follows.

> **Option value:** The *option value* associated with a possible choice is the expected value of having this choice available.

To use a standard example from the literature, it can be in your self-interest to support a tax for the maintenance of Sequoia National Park even if you are not sure that you will ever visit. The fact that you might some day choose to visit the park, and can expect the visit to be worthwhile if you do, gives this possible choice option value.[3]

The concept of option value is frequently applied within the effective altruism movement. For example, the Open Philanthropy Project has argued that working on many different causes gives it the option value of being able to focus on any of these causes in the future.

Now suppose that one is considering leaving some state A (e.g., being an apolitical movement) in order to enter some state B (e.g., being a political movement). While one is in state A, one derives option value from the possible choice to enter state B. One loses this option value, of course, by actually entering B. But one may also gain option value from the possible choice to re-enter A.

---

[2] Although our main focus here is on strategic choices for the effective altruism movement as a whole, it is worth noting that reversibility is often salient in the decisions made by individuals within the movement as well. For instance, the choice to leave a high status career for a lower status one may be difficult to reverse. It can also be very difficult for individuals to reverse a reputation for being a bad apple.

[3] See Richard C. Bishop's Option Value: An Exposition and Extension (1982) for a more thorough discussion of this example and the concept of option value.

The right decision in this case depends in part on how much option value one would gain. A good heuristic here is to ask how *reversible* the decision to leave state A for state B would be.

We define:

> **Reversibility:** The *reversibility* of a decision to leave state A for state B is the reciprocal (i.e. inverse) of the direct cost of returning to A.[4]

To better understand the significance of reversibility, let us consider a concrete case.

Suppose, again, that the decision to become a political movement has both a low direct cost and a very low reversibility. Then many future opportunities that will be available to political movements (such as the opportunity to partner with an influential activist group) could in practice be available to apolitical movements too, since the direct cost of becoming political is low. In contrast, future opportunities that will be available to apolitical movements (such as the opportunity to attract a wide range of recruits) will not in practice be available to political movements, since the direct cost of becoming apolitical again will be too high.

Note that reversibility is not the only determinant of option value. However, it is one particularly significant determinant, which it is easy to conceptualize and make snap judgements of.

In this article, we are focusing on decision situations where the decision to exit the *status quo* is easier to make than reverse. However, this choice of focus is not arbitrary. In many situations where we might consider two different strategies, one much easier to switch out from than the other, the question of which strategy we should follow will only be a live one if we are currently following the strategy that it is easy to switch out of. Due to this selection effect, we will less often be considering situations in which it is easier to re-enter the *status quo* than to exit it.[5]

---

[4] This definition can be made more precise. By the "direct cost" of switching states we mean the cost of switching itself, rather than the opportunity cost of no longer being in the initial state. For instance, in the case of Sequoia National Park, the direct cost of moving from a developed to an undeveloped state would include the financial cost of demolishing buildings, replanting trees, importing animals, and so on, but would not include the lost tax revenue from any businesses displaced. Note also that for cases where it is in fact impossible to switch states, the direct cost is infinite.

[5] Much of our analysis is applicable to cases where the more reversible decision is not the *status quo* as well. This includes situations where we face two new options which differ in terms of reversibility.

How to choose?

Let us now have a closer look at how to evaluate a potential decision by a social movement to switch strategies (e.g., to become explicitly political). We will first look at five fundamental considerations, before turning to eight secondary considerations.[6]

Fundamental considerations

Core expected value. We define a strategy's *core expected value* to be its expected value given that one does not switch out of it. A large core expected value can obviously trump concerns having to do with reversibility. For instance, if a large movement has much greater core expected value than a small movement, then it may be worth growing large even if the decision is not reversible.

Reversibility. Lower reversibility normally counts against a decision to adopt a new strategy (though see *uncertainty about core expected value*), since this implies that we will not be able to derive much option value from the possibility of switching back.

Direct cost. As also discussed in footnote 4, we define the *direct cost* of a decision to switch strategies as the cost of the decision itself, rather than the opportunity cost of no longer following the initial strategy. For example, the direct cost of switching from a small movement to a large movement might include time and money spent on outreach and the loss of any members who are alienated by the growing process (but not the new size of the movement itself). A large direct cost would of course count against a decision.

Uncertainty about core expected value. The reason reversibility can be so useful is that the core expected value of many decisions is very uncertain. For instance, it is highly uncertain how valuable it would be for the effective altruism movement to grow large. If the movement could reverse that strategy once it had embarked on it, uncertainty would be less of a problem.

Conversely, if we are certain of how valuable our strategies are, reversibility does not matter. But neither does it matter if we believe that there is no way for us to learn more about how valuable they are. Under such radical uncertainty, having the option to reverse your decision is of no avail.

However, the most common scenario is that of more moderate uncertainty. For instance, if we embark on a certain strategy, we tend to improve our estimates of its core expected value. If

[6] The fundamental considerations are the ones that we believe it is most important to take into account when choosing whether to make a hard-to-reverse decision. The secondary considerations are ones that we believe are either less important or useful mainly insofar as they help us to think more clearly about the fundamental considerations. However, this distinction is quite rough.

we learn that it is lower than we thought, having the option to reverse that strategy can be crucial.

Uncertainty can also diminish prior to the launch of a strategy, either automatically or as a result of research and testing (see *researching and testing strategies*). If we expect uncertainty to be reduced in the future, this can be a strong reason to delay a relatively irreversible decision to embark on a new strategy.

Uncertainty about reversibility and direct cost. In addition to being uncertain about expected value, we are also often uncertain about reversibility and direct cost. For instance, it seems very hard to assess how reversible a decision not to have norms of integrity actually is. This may be a reason against prematurely leaping onto paths which we suspect can be highly irreversible. We may want to wait until we have got a more resilient estimate of reversibility and direct cost, e.g., thanks to research or testing (cf. *researching and testing strategies*).

Secondary considerations

Risk aversion. Risk aversion is normally a reason not to make a hard-to-reverse decision. Since reversibility gives you the option to switch course if your strategy underperforms, it normally reduces the risk of a truly bad outcome. Note, though, that the standard view within the effective altruism movement seems to be that altruists should not be risk-averse.

Focus on long time horizons. If the effective altruism movement remains active for decades, and significant opportunities to do good continue to exist, then it will probably be faced with a large array of opportunities, some of which will only be available to the movement if it is pursuing a particular strategy (such as having a reputation for integrity). Longer time horizons also leads to greater uncertainty about which opportunities may eventually arise or become valuable (cf. *uncertainty about core expected value*). This means that it may be crucial to keep our options open, by avoiding hard-to-reverse decisions. On the other hand, if the effective altruism movement will be short-lived, or if the best opportunities to do good are fleeting, then option value considerations may not be very significant. Option value can be thought as a kind of capacity, which the movement may or may not take advantage of in the future.

Cause-neutrality and option value. One of the key features of effective altruism is *cause-neutrality*: the notion that we should not prejudge what cause to invest in, but rather compare all causes impartially.[7] If the movement finds new and more valuable causes in the future, it can pursue them. This gives the movement much greater option value compared to cause-partial groups, which are set on pursuing certain causes. It is not implausible to believe

[7] Thus, we use the term "cause-neutral" in the sense that one of us, Stefan, calls "cause-impartiality" in his article Understanding cause-neutrality.

that most of the effective altruism movement's expected value derives from this option value. However, some hard-to-reverse decisions may make it significantly harder to pursue some causes. For instance, turning political may make it harder to work on promoting bipartisan civility. This means that making hard-to-reverse decisions on key questions may deprive the effective altruism movement significant proportions of the value that cause-neutrality gives it.

Correlation between irreversibility and uncertainty. First, we saw that a question of what strategy to pursue on a certain issue often ceases to be a live one if we take a hard-to-reverse decision. Second, we are normally more certain of the expected value, direct costs, and reversibility of strategies that we already have pursued. Together, these two premises entail a negative correlation between reversibility and uncertainty: if a decision to pursue a certain strategy is hard to reverse, we typically have not pursued it previously, which normally means that we are uncertain about its expected value, direct costs, and reversibility.[8] If so, that can strengthen the case against hard-to-reverse decisions to adopt new strategies (cf. *uncertainty about core expected value* and *uncertainty about reversibility and direct cost*).

Side effects on other decisions. Deviations from the *status quo* on one issue are likely to affect the core expected value, reversibility, and direct cost of other decisions in ways which are hard to predict. For instance, growing the effective altruism movement may affect the core expected value, reversibility, and direct cost of the possible decision to become explicitly political in unpredictable ways. That may be a reason not to make several hard-to-reverse decisions at once.

Researching and testing strategies. Often, it is possible to learn more about the core expected value, reversibility, and direct cost of a decision to adopt a new strategy prior to embarking on it. This can be done through research, or through testing the strategy on a small scale. Whether to research and test a specific strategy depends on costs and expected information value. The information value is, in turn, dependent on estimated reversibility. Everything else being equal, a low level of reversibility is a reason to invest more resources in researching and testing a strategy prior to pursuing it.

Overconfidence. Humans tend to be biased towards overconfidence. That may make us underestimate the actual uncertainty of core expected value and reversibility and therefore the importance of reversibility. In particular, we may underestimate the number of doors that a hard-to-reverse decision closes. We often employ inside-view thinking (cf. Robin Hanson) to model plausible future scenarios in terms of specific causal pathways. When doing so, we often overestimate the extent to which the pathways we have identified exhaust the space of plausible scenarios. We often miss important ways in which the future could pan out, and in

[8] It should be said, however, that in some cases we may have firm knowledge of the value of hard-to-reverse strategies from other sources. For instance, deciding to change your career from earning to give to academia may be hard to reverse, but you can still get a fair estimate of its value through looking at other people's careers.

some of those, it may be very valuable to be in the state (e.g., being a small movement) that is difficult to re-enter into.

The unilateralist's curse. Many hard-to-reverse decisions are instances of the unilateralist's curse; a concept described in a paper by Bostrom, Douglas and Sandberg. The "curse" appears in situations where a group's decision may effectively be determined by a unilateral action from a single member of the group. For instance, if one person decides to tell the object of a surprise party in advance, they have effectively made the decision for the whole group. In particular, the curse predicts that the more members the group has, the more likely it is that the decision will be determined by unilateral action, regardless of whether it is the right one.

All of the examples discussed in the paper by Bostrom et al., such as the case of the spoiled surprise party, are also examples of hard-to-reverse decisions. However, there are also many hard-to-reverse decisions which cannot be undertaken unilaterally.[9] For instance, it is probably hard for a single effective altruist organization to grow the movement by itself, in the face of resistance from other parts of the movement.

Out of the three discussed examples, decisions regarding norms of, e.g., honesty and integrity are probably the most susceptible to the unilateralist's curse. For such decisions, low reversibility and the unilateralist's curse have compounding effects. That a small group can unilaterally make a decision which irreversibly harms the whole movement may pose a serious risk.

Bostrom et al. suggest that the unilateralist's curse can be lifted through deliberation or deference to other actors: what they call *the principle of conformity*.[10] In short, they argue, members of a group should agree to a code of conduct that makes it unlikely that any individual member will take the unilateral decision against the wishes of the group.

It is hard to say in the abstract which of these considerations are most important, but in our view, what one should look at first is *core expected value, reversibility,* and *direct cost*. Some of the secondary considerations are also quite important. These include *focus on long time horizons*, *correlation between irreversibility and uncertainty*, and *side effects on other decisions*. It could be useful to reflect on them, especially because they are less obvious than the fundamental considerations.

[9] Conversely, there are some decisions prone to the curse which are not hard to reverse. For instance, suppose that a member of a group can veto a decision to take a certain offer. Suppose also that the offer will not cease to be given. That is a unilateralist's curse situation, and yet the decision is not hard to reverse.

[10] There may be an empirical correlation between a refusal to adopt the principle of conformity, and a tendency to rashly take irreversible decisions. This could be mediated by overconfidence in one's own present judgements of the best course of action. However, this is merely a conjecture. The question should be studied further.

Where do all the social movements go?

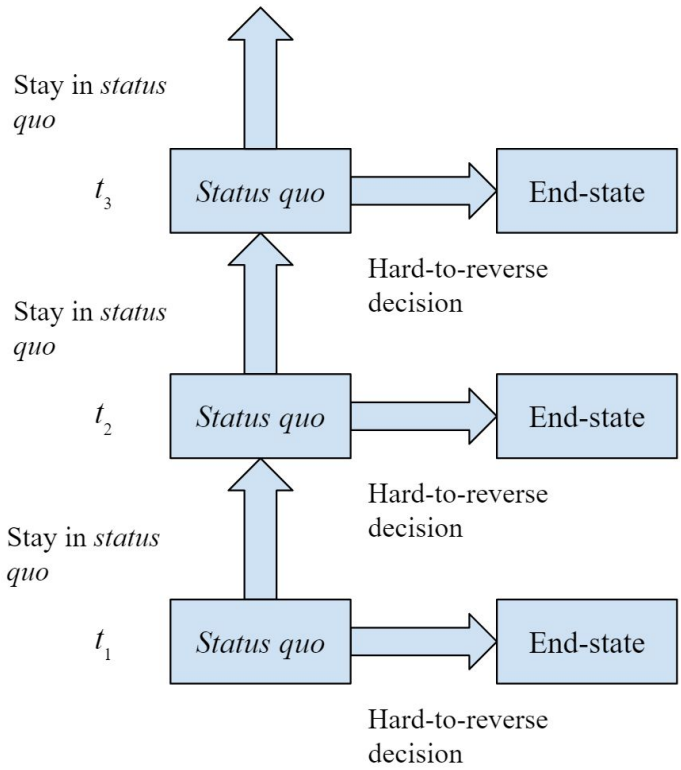
To put the concept of reversibility in perspective, let us note that it could help to explain and predict the trajectory of social movements. Suppose that:

a) A social movement has an ongoing or recurring opportunity to make a hard-to-reverse decision
b) Given that the movement survives, the probability that it takes the hard-to-reverse decision at any particular point in time never dips below some (potentially very low) lower bound.

Then:

c) If the movement exists for long enough, it will almost certainly take the hard-to-reverse decision eventually.[11]

Figure: A social movement facing a recurrent opportunity to make a hard-to-reverse decision.



This means that if it is very difficult to reverse the decisions to grow large, to go explicitly political, or to give up on norms of honesty and integrity, we may expect most social

[11] Though note Keynes's quip that "in the long run we will all be dead".

movements capable of entering these states to end up in them. In particular, we may expect that the effective altruism movement will do so by default.

This means, in turn, that the effective altruism movement should think through carefully whether those end-states are indeed desirable. That depends on host of considerations. We have addressed some of them here, but individual strategic decisions must be decided on a case-by-case basis. If the movement decides that a particular end-state is undesirable, it should reflect on what can be done to prevent us from ending up in it, e.g., through unilateral action.