Effective Altruism and Systemic Change

Antonin Broi

This is the accepted version, forthcoming in *Utilitas*. Please cite the published version.
https://doi.org/10.1017/S0953820819000153

Effective Altruism and Systemic Change

Antonin Broi

Abstract

One of the main objections against effective altruism (EA) is the so-called institutional critique, according to which the EA movement neglects interventions that affect large-scale institutions. Dietz (2018) has recently put forward an interesting version of this critique, based on a theoretical problem affecting act-utilitarianism, which he deems as potentially conclusive against effective altruism. In this paper I argue that his critique is not as promising as it seems. I then go on to propose another version of the institutional critique. In contrast to Dietz's version, it targets not the core principles of effective altruism but rather some important methodological assumptions made in EA research, namely diminishing marginal returns and low-hanging fruits. One key conclusion is that it may be time for critics of effective altruism to shift their attention from the theoretical core principles of effective altruism towards the methodological tools actually employed in practice by the EA movement.

Keywords: effective altruism, utilitarianism, cooperation, institutional change, diminishing marginal returns.

I Introduction

Effective Altruism (EA) is a movement which is about "using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis"[1]. The rapid development of the EA movement in recent years has sparked various criticisms (Gabriel, 2016). Among them is the so-called institutional critique, which has been examined several times, including within the movement[2] and, increasingly, in the academic literature. The basic claim of the critique is that effective altruism is not doing enough, or is not concerned enough, about some large-scale changes in society. For example, Herzog (2016) claims that "[Effective Altruists] seem to be missing something essential about the world in which we live: they don't look at the structures of society that are in most urgent need of transformation", and that Effective Altruists "take the current institutional order as given, implicitly denying that it can be transformed". Associated to this line of criticisms is the worry that EA is too individualistic: "The tacit assumption [in effective altruism] is that the individual, not the community, class or state, is the proper object of moral theorising" (Srinivasan, 2015).

It has proved challenging to understand these criticisms in a conclusive way[3]. At first glance, there seems to be nothing in effective altruism that rules out undertaking institutional changes, if the latter turn out to be the most effective way to do good. As a matter of fact, some interventions recommended by EA organizations, especially in animal advocacy, do clearly aim at institutional change[4]. As to the objection that effective altruism is too individualistic, it is usually met with incomprehension: Is it not the case that, as McMahan argued, "I am neither a community nor a state" and that "I can determine only what I will do, not what my community or state will do"

1 https://www.centreforeffectivealtruism.org

2 For a short overview of some responses of the EA movement to the challenge, see https://concepts.effectivealtruism.org/concepts/institutional-change

3 Another important consideration to bear in mind is that these initial institutional critiques date back to a few years ago. Given the rapid intellectual evolution of the EA movement, they might aim at an obsolete version of effective altruism.

4 For example, Animal Charity Evaluators (ACE), the main EA-aligned evaluator in animal advocacy, currently recommends "Legal and Legislative Work" as high-priority cause area. See https://animalcharityevaluators.org/advocacy-interventions/prioritizing-causes/causes-we-consider.

(McMahan, 2016)? Berkey (2018) has reviewed several versions of the objection and concluded that it either represents an implausible position or fails to contradict EA principles. However, Dietz (2018) has more recently come up with an interesting version of the critique, one that explicates its individualistic aspect by singling out the act-utilitarian principle allegedly at the core of effective altruism. This principle, he argues, implies a problem for effective altruism.

In this paper, I first show that his interpretation of the critique may not be as promising as it seems (Section II), even though he hints at an interesting feature of systemic change. Drawing on Dietz's insights, I then propose another version of the critique, which moves the focus from the foundational underpinnings of effective altruism to the methodological tools commonly used in EA research (Section III).

II Is Dietz's Institutional Critique Promising?

Dietz defends the claim that, in some cases, EA agents might fail to cooperate to achieve the most good because they adopt a principle of aiming for the most good which is overly individualistic. As institutional change arguably requires a collective endeavor based on cooperation, this means that EA agents might ignore institutional change. His critique is based on a fundamental issue concerning act-utilitarianism, one that was a topic of discussion in the seventies and to which Regan devoted an entire book (1980). Dietz directly transposes the issue from act-utilitarianism to effective altruism.

Here is the argument made by Regan and Dietz: Act-utilitarianism, and arguably one of the core EA principles (at least according to a utilitarian interpretation), hold that "[a]n act is right if and only if it has at least as good consequences under the circumstances as any other act open to the agent" (Regan, p. 12). In addition, a reasonable requirement for both act-utilitarianism and effective altruism is that it is always the case that agents who follow their recommendations collectively

achieve the most good, that is, “the class of all agents produce by their acts taken together the best consequences that they can possibly produce by any pattern of behaviour” (Regan, p. 4). If it were not always the case, then it would open the possibility that EA agents fail to collectively bring about the most good, and both act-utilitarianism and effective altruism would somehow seem deficient. This is precisely what Regan attempts to show, by appealing to a particular game-theoretic situation faced by two act-utilitarian agents, call them Alice and Bob, who have to choose between two options, A and B. If they both choose A, they get an outcome of value 2. If they both choose B, they get an outcome of value 1. If one of them chooses A and the other B, they get an outcome of value 0 (see Figure 1)[5].

		Bob	
		A	B
Alice	A	2	0
	B	0	1

Figure 1

Their decision situation, including the outcomes and their value, are objective; they do not depend on what Alice and Bob think about them. Here the outcomes of value 2 and 1 are Nash equilibria, that is, outcomes in which the option that each agent chooses, *given the option that the other agent choose*s, yields the most value, so that each agent seems to succeed in maximizing the good. Given that Bob chooses B, Alice maximizes the good by choosing B too, and vice versa. The same holds for A. Now, because there are two Nash equilibria in this game, there are two outcomes that enable each agent to individually maximize the good. The problem is that the outcome of value 1, despite being a Nash equilibrium, is clearly a suboptimal outcome, because Alice and Bob could have achieved better by both playing A and getting the outcome of value 2. This paradox motivates Regan and others to supplement the basic principle of act-utilitarianism with principles that identify

5 This game is similar to the Footballers’ Problem (studied for example by Sugden, 2003) and identical to the well-known Hi-Lo game (studied by Bacharach, 2006).

the outcome of value 2 as the only morally good outcome, and thus option A alone as the satisfactory choice for act-utilitarians. This is what Regan proposes with his cooperative utilitarianism, and Dietz with his notion of collective obligations, which he spells out in terms of team reasoning, a concept drawn from decision theory[6]. These ideas have in common that they enable the individual agent to take part in a coordinated collective action by identifying other agents who are likely to be cooperators, that is, agents who would also endorse team reasoning. These agents understand the decision situation and are ready to cooperate with her to choose A and collectively achieve the most good. Because Dietz takes one of the core principles of effective altruism to be act-utilitarianism, his line of argument seems to show that EA agents might collectively fail to achieve the most good.

I will reply to Dietz's critique in two ways[7]. First, I will argue that it is theoretically contentious. I do not have enough space to go into the details, so I will simply point to the gist of the problem. Second, and more importantly, I will argue that regardless of whether it is theoretically successful, there is no reason to think that the EA movement finds itself, or will ever find itself, in situations to which the theoretical case applies.

Dietz's argument relies on the idea that supplementing act-utilitarianism and effective altruism with further principles will prevent these situations of collective failure to achieve the most good. Indeed, to be successful, the critique against effective altruism would have to show that there are principles that are more appropriate than the mere (individualistic) act-utilitarian principle to guide the agents. That is, if the agents were to adopt these more appropriate principles, there are situations in which they would fare better than if they stuck to the individualistic principle. I will try to show that it is doubtful that such situations exist, by considering successively two ways in which an agent can find herself in a situation correctly described by Dietz's coordination game, according to

6 See Gold (2012) for an introduction to team reasoning.

7 Besides the reply that I develop in more detail, it is also interesting to note that Dietz's critique, which builds on a fundamental issue concerning act-utilitarianism and presupposes a specific conception of rationality, is unlikely to be a faithful interpretation of what critics have in mind.

whether or not she has a correct understanding of the situation and correctly identifies potential cooperators.

First, suppose that an EA agent, Alice, finds herself objectively in Dietz's decision situation, but that she is mistaken about her decision situation, or does not believe that the other agent, Bob, is an EA agent or has the correct understanding of his decision situation. In this case, for a variety of good reasons she might well end up choosing B. For example, maybe she thinks that Bob, in his mistaken appreciation of the situation, would be led to choose B. Here, it appears clearly that no collective obligation or decision procedure for advancing cooperation in addition to the act-utilitarian principle would have helped the agents choose A and achieve the outcome of value 2. If Alice does not correctly understand the situation, or does not think that Bob correctly understands the situation, she has no reason to appeal to team reasoning or to act on a collective obligation. In other words, endorsing one of these further principles in addition to the act-utilitarian principle would make no positive difference in how she acts in the coordination game. As a result, this is not an interesting case for criticizing effective altruism, because alternative principles would not have enabled the agents to achieve the best outcome.

Second, suppose that Alice does have a correct understanding of her decision situation and believes that Bob correctly understands the situation as well. This case is probably the one that we intuitively have in mind when we reflect on the coordination game. It is also the one which seems most puzzling and paradoxical: it would open the possibility that two rational agents, correctly informed about their decision situation, might nonetheless still fail to reach what they both consider the optimal outcome. This is therefore the crucial case where Dietz has to show that adding some collective obligation or team reasoning would be an improvement over the act-utilitarian principle alone. By adopting these additional principles, the agents would be led to consider the others as cooperators, and seek to undertake the action that is part of the pattern that produces the best outcome. Now, the problem is that, under these conditions, it is not clear that the act-utilitarian

principle alone is not sufficient to lead the agents to do just that. In other words, it is far from obvious that two agents abiding by the act-utilitarian principle alone, with the correct understanding of the situation and mutually thinking that the other understands the situation, could possibly be led to choose the suboptimal action B.

Whether, under these conditions, the act-utilitarian principle alone is enough to guide the agents to the optimal action is a question that traces back to some fundamental, unresolved issues in decision theory. As a first pass note that intuitively, if we were one of the players in this coordination game, we would choose option A without any doubt, and would expect, seemingly justifiably, the other agent to play the same. We could even be tempted to regard the other player as irrational if he were to choose B, as Ross (2018, Section 2.5) seems to suggest. Now, do we need collective obligations to justify these intuitions? I think that they might turn out to be superfluous. First, if we endorse evidential decision theory, it might be argued that our choosing this option is a sign that the other agent is choosing the same option, perhaps because we believe that the other agent has the same deliberative mechanism as ours (Ahmed, 2014, chapter 4). This gives us an individualistic reason to choose option A. More generally, whether or not it is true that EA agents who endorse the individualistic act-utilitarian principle alone would choose the optimal outcome all the time, Browne (2018) has argued that the conditions necessary for team reasoning to operate, and thus for the agents to choose the optimal action via team reasoning, are also sufficient to make the agents choose the optimal action on individualistic grounds alone. Indeed, it seems irrational for an agent to employ team reasoning if she knows that the other agents are not going to do so. It is only appropriate if she knows that the other agents are going to employ team reasoning. But as soon as she knows that, she no longer needs to employ team reasoning to get to choose option A. She already has the assurance that the others are likely to choose A as well, and can choose A based on an individualistic act-utilitarian principle alone. In other words, “the conditions needed to give to a team member reason to employ team reasoning make team reasoning unnecessary” (Browne, p. 13).

This would imply that there is no situation in which team reasoning makes the agents fare better than the individualistic act-utilitarian principle alone.

Overall, regardless of how we understand the coordination game, it is at least contestable that the kind of collective failure highlighted by Dietz is possible. As a result, his challenge to the individualistic act-utilitarian principle, and thus to arguably one the core EA principles, seems indecisive.

Whether the theoretical challenge to act-utilitarianism and effective altruism is sound or not, a convincing institutional critique against effective altruism would need to show that the EA movement is or could be confronted with such a situation. Otherwise, the critique would appear to miss its target: it would criticize effective altruism in general on the basis that its ethical underpinnings might in principle fail to secure some intuitive requirement. This is an interesting result for its own sake, but we should be hesitant to call it an objection against effective altruism specifically. I thus agree with Dietz that “if it were true that recognizing these [collective] obligations would not make a practical difference, this would significantly detract from the force of the institutional critique of EA”, because “EAs are primarily concerned not with purely theoretical questions but rather with how to do the most good in practice” (Dietz, p. 9). Ultimately critics would like to point to an actual example of a failure on the part of the EA movement to collectively achieve the most good as a result from its overly individualistic principle. I agree with Dietz that this is a difficult task, but unlike him I think that the task is so difficult that it makes this version of the institutional critique unpromising. To show this, I shall review several requirements for a convincing actual example of collective failure in the EA movement, and provide evidence that no real-life situation is likely to satisfy them.

What is needed is a decision situation faced by the EA movement that can be objectively described in terms of the previous coordination game. In addition, the decision situation has to be

correctly understood by the agents. Otherwise, as we have seen, no further principle can adequately supplement the act-utilitarian principle.

Dietz tentatively suggests that funding political campaigns like Clinton's during the 2016 US presidential election might qualify as an intervention that can be described in terms of the coordination game (p. 7). It would have yielded the best outcome if the entire EA movement had participated (like option A), while funding the charities highlighted by EA evaluators would be the suboptimal, individualistic intervention (like option B). The problem is that funding already highly funded political campaigns is an action that is unlikely to be understood by the agents as option A of Dietz's coordination game. Indeed, Dietz's claim is that the entire EA movement funding her campaign would have made her win the election. But, as he himself notes, it might as well be true that she would have lost even with the additional funding of the EA movement, or that only part of the additional funding would have been required for her to win. Alternatively, we could also imagine that she would have won without any additional funding. Therefore, to consider the impact of one agent funding her to be zero and that of several agents together to be relatively large, as Dietz's coordination game requires, corresponds only to one possibility among others, and the agents have no reason to think that it is the most plausible. Indeed, whatever their *actual* impact is (that is, whether or not their action actually makes her win), they might rather estimate their impact in terms of their effect on her probability of winning. The latter would be their *expected* impact, that is, their best prediction of the actual impact given all the evidence available to them. In this case, there is no reason to expect that one EA agent funding Clinton's campaign would have no impact at all on her probability of winning, while several EA agents funding her campaign would have such a high impact. Because her campaign was already highly funded, they would rather expect one agent funding her to have a small effect, and several agents to have an effect roughly proportional to their number (perhaps less than proportional if there are diminishing marginal returns, as we shall see in

the next section). They are thus unlikely to believe that they find themselves in a decision situation that corresponds to the coordination game.

Moreover, there is a sense in which Dietz's coordination game reflects a problem of coordination that is usually absent in real life. To see this, it is important to note that in real life most coordination problems seem to depend crucially on a lack of communication between the agents. Take again the action of funding a political campaign and suppose, for the sake of the argument, that this action, if undertaken by all the agents together, does yield the highest impact. Intuitively, it seems that what could prevent agents from undertaking it is that they are not in a position to communicate with each other and agree on funding the political campaign together. Now, in practical cases there is no obvious reason why an individualistic act-utilitarian agent would be any less inclined than an agent endorsing a collective obligation to communicate with other agents in order to coordinate with them. After all, in real life, efforts to reach out to like-minded agents (and thus further communication and coordination with them) are actions in their own right, which the agents can decide to undertake or not[8]. As a result, when thinking about what to do, individualistic act-utilitarian agents would assign a specific expected utility to these coordination-improving actions. Insofar as these actions might increase the probability that like-minded agents subsequently come to agree on undertaking a high-impact collective effort, it is plausible that their expected utility would exceed that of immediately undertaking an effort such as funding the charities highlighted by EA evaluators without reaching out to others beforehand. In any case, the possibility for the agents to undertake these coordination-improving actions implies that their actual decision situation does not correspond to Dietz's coordination game. We can conclude that real-life cases of coordination problems do not obviously pose a problem for individualistic act-utilitarian agents in the same way that the theoretical coordination game allegedly does. This is consistent with the observation that there have always been many opportunities for communication and coordination

8 This stands in sharp contrast to Dietz's coordination game, where efforts to communicate are apparently not considered as actions in the same way that A and B are.

within the EA movement, which might be difficult to reconcile with Dietz's individualistic picture of the EA movement[9]. They may plausibly be interpreted as the result of many individualistic act-utilitarian agents recognizing that by reaching out to other agents with broadly similar values they can increase their positive impact. There is no need to appeal to collective obligations to account for this observation. So, if EA agents were considering interventions that are optimal only if undertaken collectively, they would most likely communicate and coordinate, and this on individualistic grounds alone, provided they have good reasons to think that other EA agents have the same understanding of the situation.

Finally, this latter condition is substantial. Even if we could come up with a decision situation faced by EA agents that could objectively be described in terms of the previous coordination game, nothing guarantees that all EA agents would not only have a correct understanding of the situation, but also believe that the others have the same understanding. Of course, members of the EA movement share many values and empirical beliefs, but there are still significant disagreements. This is all the more relevant when it comes to institutional (and in particular political) change, whose effects are particularly difficult to predict and for which there is little agreed-upon evidence. While it is plausible that some subset of the EA movement has enough in common to be considered as a collective agent, it is much less so for the entire EA movement.

Overall, we have some reasons to believe that the requirements for an actual example of collective failure based on Dietz's theoretical objection are particularly demanding, so our prospects for establishing that the EA movement is collectively failing are rather bleak. Dietz concludes at the end of his paper: "In order to resist the institutional critique as I have presented it, EAs will have to defend substantial views in moral psychology and the theory of collective action" (Dietz, p. 10). I

9 In addition to frequent online activity, for example on social media or on the EA Forum (https://forum.effectivealtruism.org), real-life events like the EA Global conference are good opportunities for cooperation and sharing of ideas.

disagree. While his critique does trace back to unresolved debates in decision theory, there is no need to look into the latter to show that the critique may not be as promising as it seems to be.

Like the objections against effective altruism addressed by Berkey (2018), Dietz's version of the institutional critique targets core EA principles. The focus on these principles is understandable. They are indeed *philosophical* principles, and are widely regarded as the essence of what the EA movement is about. However, in this paper I would like to point to another direction for research. Rather than the core EA principles, there seems to be room for criticism in the set of secondary assumptions, heuristics and methods that are used in EA-aligned research. The version of the institutional critique that I propose is a first step in that direction, and a first attempt towards exploring how EA principles are implemented in practice.

III Towards a Methodological Institutional Critique

The argument made by Dietz hints at an interesting feature of systemic change, which he was arguably going after in his example of funding political campaigns. Suppose that options A and B in Dietz's coordination game represent possible actions that EA agents could undertake. In this game, if one unit of resource, i.e. the efforts of one agent, is put into an action, no positive impact is brought about (the value of the outcome is zero). However, two units of resources yield a positive impact. In other words, these actions bring about positive impact only if several agents undertake them. This is thus a special case of *increasing marginal returns*, the assumption that the marginal impact of a unit of resources spent on an action increases with the amount of resources already spent on the action. In Dietz's coordination game, the resources that are spent have no impact until a certain threshold is reached, after which they have a significant impact, arguably through systemic, or institutional, change[10]. Increasing marginal returns, especially in the case of a threshold, are

10 In this paper I consider "systemic change" and "institutional change" as interchangeable.

commonly associated with actions that affect large-scale political, social and economic institutions. Indeed, institutions are difficult to disrupt, but when they are, after a critical mass of effort is exerted on them, the disruption can potentially bring about very positive change and enable to reach an improved equilibrium in the system. Intuitively, there are increasing marginal returns at the threshold point where the critical mass of efforts bears its fruit and starts to bring about systemic change. Because the last units of resources yield much more impact than the previous ones, these actions are likely to be worth undertaking only if a large amount of resources can be spent on them. Otherwise, they are unlikely to be the best way to allocate our scarce resources.

I will focus on this feature to construct a new version of the institutional critique, as it seems to go a long way in explaining the emphasis that many critics put on collective agency, as well as their rejection of individualism[11]. This feature also helps to make sense of the idea that by considering what the efforts of many agents can produce together, we really open the possibility of effective collective actions. Before I go on, a few remarks are in order.

Talk of increasing marginal returns requires the existence of a returns function, which maps each amount of resources devoted to the action to a value representing the quantity of good done, that is, its positive impact. The amount of resources, or cost, can be quantified in terms of e.g. money or working time, and the impact in terms of a common currency such as the number of lives saved. For example, we can vary the amount of money that we spend on distributing bednets to people in developing countries to prevent malaria: a number of lives saved can be associated with each amount of money spent on this intervention. Returns functions may apply to resources spent on *interventions*, like distributing bednets, directly transferring money to the poorest, implementing

11 The critique that I am going to develop here could be made in a formal way, by using a mathematical framework of cost-benefit analysis of interventions. While this would have the advantage of being systematic and providing well-defined necessary and sufficient conditions for the problems I am going to discuss, it would exceed the scope of this paper. Here I will simply offer an intuitive grasp of an idea that may be promising to investigate further.

deworming programs, or lobbying for trade reform[12], or to resources spent on *cause areas*, like global poverty, industrial animal farming, etc.

Let us call *systemic interventions* those interventions that have increasing marginal returns for some large amounts of resources spent on them[13]. In defining systemic interventions in this way, I set aside other putative features that could be associated with interventions that affect systems, like high probability of failure[14], scale of effects (e.g. large number of people affected), or importance of long-term effects. The definition also leaves open the possibility that other interventions than those affecting social systems have increasing marginal returns, and thus are systemic interventions in the sense employed here. However, most of the intuitive examples I will draw on have to do with social systems.

It is difficult to come up with uncontroversial examples of systemic interventions in this sense. I have already shown that funding highly funded political campaigns is unlikely to display increasing marginal returns (see Section I). Critics of effective altruism might point to other actions that purport to change global political or economic institutions (advocacy in general, including campaigning, lobbying, etc.) or social or cultural norms (through media campaigns, education, individual persuasion, etc.) but it is controversial whether these actions really feature increasing marginal returns at some point. Of course, insofar as they deal with systemic changes that are all-or-nothing (for example, either a law is passed or it is not) and that we can make efforts to push for these changes, it is likely that there be a threshold of resources that is enough to trigger them. In this case, it is true that the last unit of resource that *actually* enables to reach the threshold does yield increasing marginal returns, but this does not mean that the *expected* returns function also has

12 Apart from the last one, these are all examples of interventions that GiveWell, the main EA-aligned evaluator working on global poverty, recommends at the present time.

13 To define precisely what is meant by “large” would unfortunately take me too far. Roughly speaking, it is to be understood as large enough to trigger systemic change.

14 Interventions that have a high probability of failing are at the core of OpenPhilanthropy’s hits-based conception of philanthropy (See https://www.openphilanthropy.org/blog/hits-based-giving). Funding Clinton’s campaign is an example of intervention that was likely to fail (if Clinton loses the election), but not obviously an example of systemic intervention in my sense, as I have shown in the previous section.

increasing marginal returns, because we usually do not know where this threshold lies. Now, insofar as the actual impact and the actual returns function remain unknown, we are mostly interested in the expected impact and the expected returns function. It might be reasonable to assume that as we spend more and more efforts into attempting to trigger the systemic change, the probability of triggering the systemic change decreases. This would imply diminishing expected marginal returns.

The version of the institutional critique that I want to develop starts with the following question: If there were an effective systemic intervention available, would the EA movement undertake it? I wish to support the claim that it might not undertake it because some assumptions and heuristics widely used make it difficult to take such interventions into account. Importantly, I leave aside the question whether there indeed exists an action that is systemic, available to the EA movement, and effective enough to be the best option in order to achieve the most good. My goal is only to show that if there were one, it is unclear that the EA movement would undertake it. If my argument is convincing, the EA movement might currently lack the tools to engage with debates where systemic change plays an important role, for example around anti-capitalism or abolitionism in animal advocacy.

Unsurprisingly, the main assumption in the EA literature that could conflict with the existence of systemic interventions is that of *diminishing marginal returns.* That cause areas have approximately diminishing marginal returns, at least after a significant amount of resources have been invested in them[15], is a widespread assumption in the EA literature. It can be found in various influential texts of the EA movement. Robert Wiblin, in one of the main introductory texts to cause prioritization, states that "[a]fter a large amount of resources have been dedicated to a problem, you'll hit *diminishing [marginal] returns*" (Wiblin, 2017). In addition to money, the claim is considered to hold true for investments in human resources. While choosing one's career, we should remember

15 Of course, it is not clear what is meant by "significant", but the assumption is usually thought to apply to most prominent cause areas at this time, in particular global poverty.

that “[t]he more effort that’s already going into a problem, the harder it is for *you* to be successful and make a meaningful contribution. This is due to *diminishing returns*.” (Todd, 2017).

The assumption is so important that William MacAskill, one of the intellectual leaders of the movement, mentions it in passing in a recent eleven-minute TED talk, where he reviews what it takes for a problem to be worth addressing: “More neglected is better, because of diminishing returns. The more resources that have already been invested into solving a problem, the harder it will be to make additional progress”[16]. In this talk, as in many other texts, the assumption of diminishing marginal returns is closely related to one of the three factors found in the main framework of cause prioritization, the Importance-Neglectedness-Tractability framework. According to this framework, a cause area is promising if it scores well on three factors: how big the problem is (Importance, or Scale), how neglected it is by other agents (Neglectedness), and how easy it is to make headway in solving the problem (Tractability). The Neglectedness factor is often informally interpreted in the following way: the less resources are currently spent on a cause area, the more promising it is. Justifying this claim requires postulating a version of diminishing marginal returns for cause areas. In his book *Doing Good Better,* MacAskill thus affirms:

> The law of diminishing returns provides a useful rule of thumb for comparing causes. If a specific area has already received a great deal of funding and attention, then we should expect it to be difficult for us to do a lot of good by devoting additional resources to that area. In contrast, within causes that are comparatively neglected, the most effective opportunities for doing good have probably not been taken. (Chapter 4, 2015)

Why is it that the so-called law of diminishing returns generally holds for cause areas? One important justification seems to be that when focusing on a cause area, we can choose among various interventions *which have themselves diminishing marginal returns*. Initial resources spent on them have a high marginal impact, but it rapidly decreases as more resources flow towards them. As a result, when an intervention has received a lot of funding, it becomes less effective to spend

16 https://www.ted.com/talks/will_macaskill_how_can_we_do_the_most_good_for_the_world?

more resources on it than on other interventions. Those interventions which are initially very effective, but with rapidly diminishing marginal returns, are often called *low-hanging fruits*. In an influential talk, whose transcription features in the EA handbook published by the Centre for Effective Altruism[17], Cotton-Barratt uses the metaphor of prospecting for gold to introduce cause prioritization. Gold is to be understood here as the good, or value, that EA agents are seeking to achieve. The idea of low-hanging fruits is explained as follows:

> If you first go to an area where nobody has been before, then the seams of gold that are running through the ground have often been eroded a little bit, and you can have little nuggets of gold just lying around on the ground, and it's extremely easy to get gold. So you have some people go in, they do this for a bit, and they run out of all the gold on the ground. (p. 25, 2016)

The idea is quite appealing, and seems to apply to a vast range of areas. Later on in the talk, Cotton-Barratt mentions the following example in global poverty:

> I understand that 15 or 20 years ago, mass vaccinations were extremely cost-effective and probably the best thing to be doing. Then the Gates Foundation has come in and funded a lot of the mass vaccination interventions. Now, the most cost-effective intervention is less cost-effective than mass vaccinations [were 20 years ago]. That is great because we have taken those low hanging fruit. (p. 26, 2016)

By definition, systemic interventions are not low-hanging fruits. Postulating diminishing marginal returns for cause areas, based on the above considerations about low-hanging fruits, is thus likely to contradict the existence of systemic interventions. In addition to this theoretical contradiction between diminishing marginal returns and systemic interventions, tensions arise when it comes to acting on these ideas. Indeed, in practice, the ideas of diminishing marginal returns and low-hanging fruits imply that EA agents should simply fill the needs of the most effective interventions, up to the point where their marginal impact decreases so much that it equals that of the next most effective interventions, then fill the needs of the latter, and so on. This holds regardless of the amount of resources that we want to spend on the cause area. In other words, the best allocation

17 https://www.effectivealtruism.org/handbook/

always starts with the low-hanging fruits, and then goes from there to less effective interventions, so agents should always "go for the low-hanging fruits". This seems to be implicit in Cotton-Barratt's talk, as he assumes that it is best to first pick up the gold on the ground. The problem is that if systemic interventions exist, this strategy might be suboptimal: it might sometimes be worth spending all of our resources on bringing about systemic change rather than picking the low-hanging fruits first. The alternative I am pointing at here, in terms of Cotton-Barratt's metaphor, is that there might be a large seam of gold buried deep in the ground that requires a lot of human resources to reach. Now, if it is large enough, it might be better not to spend any time picking up the easy gold, and go instead for the larger seam.

Two repercussions follow from the practical recommendation of aiming for the low-hanging fruits. First, when a constant flux of resources is progressively injected into a cause area, going for the low-hanging fruits implies that there is no need to accumulate resources over time: as we receive the additional resources, we can immediately spend them on the low-hanging fruits. By contrast, if there are systemic interventions, it might be better to wait until we have enough resources in order to consider spending them on systemic interventions and trigger systemic change, rather than spend the initial flux of resources on the low-hanging fruits. Second, and more importantly, when different EA agents reflect on where to spend their respective resources on, going for the low-hanging fruits implies that there is no need for coordination, other than ensuring that the EA agents do not simultaneously spend too many resources on one low-hanging fruit whose marginal returns are rapidly diminishing. However, if there are systemic interventions, coordination between agents may be required to assess the total amount of resources that all agents together have available, and agree to spend it on a single systemic intervention to trigger systemic change.

This opens a new way of understanding the kind of collective failure that Dietz aimed at capturing with his coordination game. Imagine a variant of Dietz's game: this time, if one of the two agents chooses A and the other B, the outcome is 0.6, and not 0, as it was previously (see Figure 2

below). The decision situation can be now interpreted as involving an intervention with diminishing marginal returns, B, and a systemic intervention, A. Because of the diminishing marginal returns of B, while one agent choosing B brings about 0.6, two agents choosing B bring about only 1. The systemic intervention, on the other hand, yields positive impact and systemic change only if both agents undertake it, so it requires more resources than either agent possesses. Only if they combine their efforts by both undertaking A can they achieve the outcome of value 2.

		Bob	
		A	B
Alice	A	2	0.6
	B	0.6	1

Figure 2

My version of the institutional critique provides a new way of understanding why EA agents in this situation would not achieve systemic change. In my version, it is not the individualistic, act-utilitarian principle that Alice and Bob endorse as EA agents that is to blame, but rather their assessment of the systemic intervention and their decision procedure. Indeed, they are unlikely to correctly assess the value of the outcome that involves systemic change because it contradicts the assumption of diminishing marginal returns. Alternately, they may ignore it altogether because the decision procedure associated with aiming for the low-hanging fruits does not require considering what can be done with more resources than are available to the agent. As a result, the problem with effective altruism would not consist in the absence of coordination to achieve the optimal outcome. Rather, the absence of coordination is merely a consequence from another problem, i.e. the failure of EA agents to deal with systemic interventions.

To sum up, I have tried to point to some EA assumptions, namely low-hanging fruits and diminishing marginal returns, which may rule out the existence of systemic interventions. For this reason, if there were systemic interventions, it might be that the EA movement ignores them, or does not undertake them even though they are part of the best allocation of resources. As a result, effective altruism would fail by its own standard: the EA movement seeks to achieve the most good, but uses some assumptions that could rule out the most effective action to do so.

The version of the institutional critique developed here could shed new light on other versions of the critique. For example, the problem of competition between NGOs and states, which has been brought up several times as an argument against effective altruism (see for example Clough, 2015), could be re-interpreted as involving an alleged conflict between the low-hanging fruits of directly helping the poor through the action of NGOs and the systemic interventions associated with advocacy and empowerment of the poor. Because these interventions need to reach a certain critical mass before they bring about permanent and large-scale change in the action of the state, they would feature increasing marginal returns. The distinction between low-hanging fruits and systemic interventions can also be applied to the classic opposition between reform and revolution developed in Marxist thought (Luxemburg, 2006), which may underlie some leftist objections against effective altruism (for example, Snow, 2015, drawing on Gomberg, 2002). Reforms can improve the living conditions of the poor and can be relatively rapidly obtained. They are low-hanging fruits. However, it could be argued that only radical political change can solve the whole problem, i.e. capitalism, which is not addressed by the EA movement. Finally, in animal advocacy, some proponents of abolitionism, the view that any exploitation of animals for one's own consumption is morally bad, hold that the most efficient way to achieve the end of animal exploitation is to promote veganism (Francione, 1996), which might be understood as an effort to trigger a systemic change in

social norms. The low-hanging fruits, on the other hand, might be the "welfarist" measures to increase the well-being of animals which have been widely embraced by the EA movement[18].

Once they are understood in this way, these critiques enable to highlight an intriguing effect that picking the low-hanging fruits may have on systemic interventions. Implicit in what has been said up to now is the assumption that the interventions do not interfere with each other. Spending resources on an intervention has no effect on the returns function of other interventions. Although convenient, this assumption may turn out to be false, especially when it comes to low-hanging fruits and systemic interventions. In particular, there may be cases in which picking the low-hanging fruits to partially solve a problem might raise the cost of completely solving the problem later on via systemic change. This is what Kissel (2007) seems to be hinting at when he discusses the consequences from NGOs providing some service instead of the state:

> [T]his move can have harmful indirect effects such as when the most discerning poor people, who also were most likely to lobby the state and monitor implementation, move to the NGO service and shift their advocacy to the NGO. […]
>
> [T]he discerning poor are less likely to come together and throw out their incompetent governments when their advocacy target shifts from the state to the NGO. This reduction in political agitation makes less likely the formation of political movement capable of wide-ranging and self-sustaining solutions. (p. 18)

Here, picking the low-hanging fruits (funding the NGO) makes the systemic political change more difficult to achieve, because people become less prone to target the state for advocacy. The same logic applies to larger scale changes: revolution is made more difficult to achieve by previous mild reforms. This is because radical political change occurs following a breaking point, and the low-

18 For example, corporate campaigns for cage-free eggs have been strongly supported by EA organizations like OpenPhilanthropy or ACE. The interpretation proposed here presupposes that the abolitionist strategy may naturally be associated with (radical) systemic change, and (neo)-welfarist strategies with non-systemic, or less systemic, change. While there is some reason to do so (abolitionism does purport to change the property status of animals, which is a radical cultural and legal change), it should be pointed out that it is debatable to categorize abolitionist and (neo)-welfarist strategies in this way. Indeed, on the one hand, (neo)-welfarist strategies are mostly about corporate outreach and policy influencing, which may count as fairly systemic actions; on the other hand, Francione tends to deny the systemic status of abolitionism, as he promotes "*incremental eradication of the property status of animals"* (Francione, 1996, p. 4).

hanging fruits may prevent us from getting to this breaking point. Similarly, in animal advocacy, Francione (1996) holds that incremental progress in animal well-being might in the end hinder the likelihood of ending animal exploitation, by focusing the attention on preventable animal suffering rather than animal rights.

If going first for the low-hanging fruits may undermine the efficiency of systemic interventions in the future, it is all the more important to take systemic interventions into account when prioritizing causes and interventions, and to engage thoroughly with critics who defend them. Indeed, whether or not anti-capitalism or the abolitionist stance in animal advocacy are plausible on their own, it is remarkable that, if the version of institutional critique developed here is promising, the EA movement might struggle to engage with these positions because some methodological assumptions commonly made in the EA movement seem to contradict their existence. This cannot be a good thing, since effective altruism is all about finding out what the best interventions and strategies are, and thus should not rule out a strategy on a priori grounds.

3 Conclusion

In this paper I have attempted to provide a new direction for investigating the institutional critique against effective altruism. I first cast doubt on the version of the critique raised by Dietz, by establishing that it is not only theoretically questionable, but also difficult to apply to an actual situation faced by the EA movement. To develop a new institutional critique, I then focused my attention on pervasive ideas found in the methodology of EA research which may conflict with the kind of systemic change that EA critics seem to have in mind. This might render the EA movement oblivious to some systemic interventions, thus exposing it to justified suspicion. It also sheds new light on other objections raised against effective altruism.

More work is needed in order to develop a full-fledged institutional critique based on the outline I have presented. I have looked into assumptions that can be found in the EA literature and that seem to be taken for granted by many. For the critique to be really convincing though, it remains to be shown how these assumptions permeate the practice of EA-aligned researchers and practitioners in cause prioritization. It might turn out that the assumptions of low-hanging fruits and diminishing marginal returns do not play a significant role in investigating effective interventions after all, or play a role only in some cause areas, and not in others. Moreover, the argument developed here could be partially deflected by maintaining that ruling out institutional interventions is fine, by arguing either that they do not exist or that they are so rare that the EA movement does not miss out on important opportunities by ruling them out. But these are substantial claims which need to be discussed carefully. While more research is necessary to assess the validity of the critique put forward here, I hope to have at least offered some common ground for future debate between critics and partisans of effective altruism.

References

Ahmed, A. (2014). *Evidence, decision, and causality*. United Kingdom: Cambridge University Press.

Bacharach, M. (2006). *Beyond individual choice: teams and frames in game theory*. (N. Gold & R. Sugden, Eds.). Princeton, N.J: Princeton University Press.

Berkey, B. (2018). The Institutional Critique of Effective Altruism. *Utilitas*, *30*(02), 143–171.

Browne, K. (2018). Why Should We Team Reason? *Economics and Philosophy*, *34*(02), 185–198.

Clough, Emily. "Effective Altruism's Political Blind Spot." *Boston Review*, July 14, 2015. https://bostonreview.net/world/emily-clough-effective-altruism-ngos.

Cotton-Barratt, O. (2016). Prospecting for Gold. In *Effective Altruism Handbook* (2nd edition). Centre for Effective Altruism. Retrieved from https://www.effectivealtruism.org/handbook

Dietz, A. (2018). Effective Altruism and Collective Obligations. *Utilitas*, 1–10.

Francione, G. L. (1996). *Rain without thunder: the ideology of the animal rights movement*. Philadelphia, Pa: Temple University Press.

Gabriel, I. (2016). Effective Altruism and its Critics. *Journal of Applied Philosophy*, 457–473.

Gold, N. (2012). Team reasoning, framing and cooperation. In S. Okasha & K. G. Binmore (Eds.), Evolution and rationality: decisions, co-operation and strategic behaviour. Cambridge: Cambridge University Press.

Gomberg, P. (2002). The Fallacy of Philanthropy. *Canadian Journal of Philosophy*, *32*(1), 29–65.

Herzog, Lisa. "Can 'effective Altruism' Really Change the World?" *openDemocracy*, February 21, 2016. https://www.opendemocracy.net/transformation/lisa-herzog/can-effective-altruism-really-change-world.

Kissel, J. (2017). Effective Altruism and Anti-Capitalism: An Attempt at Reconciliation. *Essays in Philosophy*, *18*(1).

Luxemburg, R. (2006). *Reform or revolution*. New York: Pathfinder.

MacAskill, W. (2015). *Doing good better: How effective altruism can help you make a difference*. New York, N.Y: Gotham Books.

McMahan, J. (2016). Philosophical Critiques of Effective Altruism. *The Philosopher's Magazine*, *73*.

Ross, D. (2018). Game Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University.

Snow, M. (2015, August 25). Against Charity. Retrieved 17 April 2016, from https://www.jacobinmag.com/2015/08/peter-singer-charity-effective-altruism

Srinivasan, A. (2015, September 24). Stop the Robot Apocalypse. *London Review of Books*, pp. 3–6.

Sugden, R. (2003). The Logic of Team Reasoning. *Philosophical Explorations*, *6*(3), 165–181.

Todd, B. (2017). Want to do good? Here's how to choose an area to focus on. Retrieved from https://80000hours.org/career-guide/most-pressing-problems

Wiblin, R. (2017). How to compare different global problems in terms of impact. Retrieved from https://80000hours.org/articles/problem-framework