Economic Design for Effective Altruism*

Dominik Peters
Department of Computer Science
University of Oxford, UK

Early Draft – October 2017

Written for *Future of Economic Design*
(eds. Jean-François Laslier, Hervé Moulin, Remzi Sanver, William S. Zwicker)

Abstract

A growing movement of 'effective altruists' are trying to find the best ways of helping others as much as possible. The fields of mechanism design and social choice theory have the potential to provide tools allowing this community to make better decisions. In this article, I consider several avenues for such research.

Effective Altruism. Effective Altruism is the project of using evidence and reason to figure out how to benefit others as much as possible, and taking action on that basis (MacAskill, 2017). Thousands of people and many organisations now identify as Effective Altruists ("EAs"), and they have formed local groups in dozens of cities around the world and regularly meet at the Effective Altruism Global conferences. What do EAs do? Most of them donate a significant fraction of their annual income to charities that they view as particularly effective. Indeed, an early precursor to the EA movement was the Giving What We Can pledge to give at least 10% of one's income to effective charities; more than 3,000 people have now taken the pledge. EAs also invest significant effort into identifying which career path will best allow them to do the most good. But the defining and perhaps most challenging task of EAs is "cause prioritisation", that is, figuring out which actions (or jobs, or charities) actually do the most good. This usually involves designing cost-effectiveness models of various options, and gathering scientific research to inform those questions.

Effective Altruists tend to be most excited to tackle problems that have large scale (they severely affect many people), that are tractable (progress on solving the problems is possible), and that are neglected (marginal returns of working on the problems are high). Several broad fields of actions are commonly agreed to be promising on these

[*] This is an early draft. I would be very happy to hear any feedback, and in particular pointers to the literature I have missed so far. Email: dominik.peters@cs.ox.ac.uk

criteria. For example, much attention is directed towards helping people in extreme poverty through interventions such as direct cash transfers or distributing malaria nets; towards farm animal welfare through lobbying corporations or funding research into meat replacements; and towards helping future generations through funding research into mitigating global catastrophic risks (such as biosecurity, runaway global warming, or risks from transformative AI).

As interest in Effective Altruism has grown since the inception of the movement in around 2011, there has been an increasing emphasis to optimise the impact of the whole *community*, and not just the impact of each individual separately. There are possible gains from cooperation, from compromise in case of disagreement, and from identifying individuals' comparative advantage within the community.

In this article, I will outline several ways in which techniques from mechanism design and social choice theory could be used to develop tools that could help the Effective Altruism community make better decisions. For example, I will point out that there are some difficult coordination problems in deciding who should donate to what charities. Given the significant donation volume within the EA community, it is important to develop tools and conceptual insights that allow us to identify a good way of deciding where donations should go. I also outline several research questions in handling the aggregation of different moral views, which has applications in handling *moral uncertainty* and cases where a group of agents designs a systems making moral decisions (such as advanced AI system). In closing, I mention two other directions: facilitating *moral trade* and good *job allocations.*

The area of Economic Design has long studied ways in which we can improve society, make people better off, and implement outcomes that are socially optimal. Kidney exchange is a notable success story of this field, where matching theory and mechanism design have saved and improved hundreds of lives (Roth et al., 2004). Similarly, the Effective Altruism movement has the ambitious goal of improving the lives of as many people as possible. Helping EAs achieve these goals provides an exciting opportunity to apply our methods of economic design for social good.

Donation Allocation. Most effective altruists give a significant amount of money to charitable causes each year. At the time of writing, *Giving What We Can* has more than 3,000 members who have pledged to donate at least 10% of their annual income to effective charities; in 2015, they gave about $7 million (Centre for Effective Altruism, 2016). There are also some larger organisations, notably the foundation *Good Ventures*, which in 2016 granted more than $100 million to effective causes through the Open Philanthropy Project.

It has long been recognised in the community that charities have a limited *room for more funding* (RFMF): any given organisation can only hire a limited number of staff and can only scale up a limited amount without diluting its cost-effectiveness, at least in the short run. Donations to a given organisation have decreasing marginal returns. For example, the charity evaluator GiveWell (2016) estimates that its top recommended charity, the Against Malaria Foundation (AMF), could productively use between $78 million and \$191 million of funding for the year 2017.

Karnofsky (2014) noticed that this gives rise to a problem of *donor coordination*: Following GiveWell's analysis, many donors think that AMF is a top giving opportunity, and collectively, these donors would be willing to donate more to AMF than its room for more funding. Not all of these donors should give to AMF, since not all the money could productively be used. Thus, some donors should give to their second-most preferred charities.

However, donors may disagree about which cause is the best cause after AMF. In particular, they may prefer *other donors* to fill AMF's RFMF, while they donate to *their* next-best option. During the 'giving season' around Christmas time, this leads to donors waiting until the last minute to make their donation decision, only giving to a cause if it seems like other donors have not yet filled a relevant RFMF. This leads to inefficiencies (Todd, 2016).

Instead, one could imagine a centralised mechanism, a *donation clearinghouse*, to which donors communicate their preferences over charitable organisations and their donation budget, and which then decides how these budgets are allocated to different charities.

Let us consider an example. Suppose there are three charities, A, B, and C, each with a RFMF of \$1. Two donors, 1 and 2, each wish to donate \$1. They have different preferences over causes:

$$A \succ_1 B \succ_1 C \quad \text{and} \quad A \succ_2 C \succ_2 B.$$

Thus, they agree that A is the best giving opportunity, but they disagree about the merits of B and C. In an uncoordinated outcome, it could happen that both donors donate \$1 to A, which wastes \$1. (This type of outcome tends to occur in the aftermath of major catastrophes.) Or both donors could guess that the other donor will give to A, so that the money goes to B and C only; using the money suboptimallly.

The "fair" outcome, at least if we only have these ordinal preferences available, would seem to be that both donors give \$0.50 each to A, while donor 1 gives the remaining \$0.50 to B, and donor 2 gives the remaining \$0.50 to C. Note that, when using this type of reasoning, the donors may have some incentive to be strategic. For example, donor 1 could manipulate and report their preferences to be $B \succ_1' A \succ_1' C$, whence we might recommend that donor 1 gives \$1 to B while donor 2 gives \$1 to A; donor 1 would likely prefer this outcome.

The research challenge is to design mechanisms that produce high-quality donation allocations and, if possible, prevent strategic behaviour. Addressing this challenge requires answering several questions.

- How should donors report their preferences? Many users would not be happy with a system that only elicits an ordinal ranking over charities, like in our example. They might prefer a system that uses cardinal information, for example by asking for a cost-effectiveness estimate in units such as quality-adjusted life years produced per \$ donated. They may also want to input an individual judgement about the size of a charity's RFMF; and they may also wish to report preferences about how much *others* donate to various charities.

- What optimisation objective should the system use? There are many possible choices, and it is not obvious which one is the best. For example, if the system receives cost-effectiveness estimates from the users, we could first aggregate these estimates (e.g., by taking an average to exploit the "wisdom of the crowds"), identify the best charities, and allocate money only to them, subject to their RFMF. However, this system could lead to the contributions of a donor being used for a cause that they strongly disagree with.

  Alternatively, we could define the *individual welfare* of a donor by how much impact their individual donations will have, according to their own estimates. For example, if a donor i estimates charity A to have an impact of 10 'utils' per dollar, and their contribution of \$1 is allocated to go to A, then i would have a welfare of 10 utils. The mechanism could then aim to maximise the utilitarian social welfare, i.e., the sum of each donor's individual welfare.

  The latter choice of objective limits the amount of aggregation that goes on, and in particular it ignores the public-good-type character of donations: a donor will typically not only care about how much good their individual donation did, but how much good was done by all donations together. Thus, there seems to be a trade-off between incorporating the (positive) externalities of a donation, and ensuring that i's donation is only used for causes that i agrees with.

- We have already seen that some systems of this type will be manipulable by voters who pretend not to like some charities whose RFMF will likely be met by other donors. Are there mechanisms that can avoid this type of strategic behaviour? A natural starting point would be VCG mechanisms. However, to guarantee strategyproofness, these mechanisms will typically require burning money. In the context of donations, this seems unacceptable. It would be interesting to see whether there are mechanisms that are (approximately) budget-balanced, i.e., ones that do not burn (much) money.

A good mechanism should also be able to accommodate donors with very different budget sizes (Muehlhauser, 2017). For example, within the EA community, a significant fraction of donations currently come from Good Ventures via the Open Philanthropy Project, accounting for about \$100 million per year. This contrasts with donations by individuals, who typically donate less than \$10,000 per year each.

A paper by Conitzer and Sandholm (2011) considers several of these problems, viewing donor coordination as a problem of *negotiation*. In their setting, donors can propose (potentially complicated) *matching offers*, which they can use to incentivise other agents to donate to certain charities. Conitzer and Sandholm (2011) study the computational complexity of clearing these markets, finding mostly hardness results, though they present integer linear programming formulations that should be efficient in practice. They also begin a study of the mechanism design problem, obtaining some impossibility results. In my view, their precise model has some undesirable features that would make it unsuitable for use by the EA community: in particular, the model allows an agent to threaten not to donate money unless others donate to that agent's preferred choice, and often, the threat would be successful in optimum; I would want to aim for a mechanism that encourages cooperative behaviour. Still, their results are a good starting point for further study.

Moral Uncertainty. To most effectively help others, one must be able to decide which of two actions better fulfils this goal. But this question is difficult to answer, because such comparisons involve many trade-offs. If some action improves the lives of many beings but violates some rights, is this worth it? If some action improves the lives of future generations but hurts those currently alive, is this worth it? How much do non-human animals count? Is it more important to prevent suffering, or more important to increase the number of happy beings?

Different moral theories give different answers to these questions. Even after much reflection, many people find that they are uncertain about the correct answers. They have *moral uncertainty.* This uncertainty may be reflected by the decision-maker having credences (subjective probability judgements) about various moral theories being "correct".[1] For example, someone might spread 60% probability mass among various utilitarian theories, and allocate the rest among deontological (rights-based) theories and virtue ethics. Given these credences, the decision-maker asks the question: What is the morally right action for me to take?

If we formalise different theories as rankings of a set of feasible actions in order of "choice-worthiness", then the problem of selecting the right action is formally identical to the standard voting setting of choosing a winning alternative given a (fractional) profile of preference orderings (MacAskill, 2016). Thus, we may evaluate the appropriateness of standard voting rules for this setting. The standard tool for this is to employ *axiomatic analysis*, which identifies desirable properties of aggregation mechanisms and asks which rules satisfy them. MacAskill (2016) argues that, for the setting of moral uncertainty, the voting rule chosen should satisfy the *participation axiom* (Moulin, 1988), which requires that if we increase our credence in a theory that recommends the currently recommended alternative, the recommendation does not change. He further argues that strategyproofness is not a concern, because theories are not strategic actors. Thus, he proposes to use *Borda's rule*,[2] which is arguably the best of the known rules satisfying participation. In certain models, Borda's rule also maximises the likelihood of picking the best option according to a ground truth (Young, 1988).

There is potential for more work in social choice theory analysing this setting. For the approach based on maximum likelihood estimation, it would be useful to identify noise models that are plausible for the case of moral theories (for a survey of noise models studied so far, see Elkind and Slinko, 2016). Another direction is considering aggregation of non-ordinal theories. Indeed, many moral theories are not adequately formalised as ordinal rankings. While plausibly, as MacAskill (2016) argues, *some* theories are 'merely ordinal', other theories encode their prescriptions in other forms: Most flavours of utilitarianism come with a natural cardinal structure; deontic theories only specify

[1] For a moral realist, this may refer to uncertainty over which moral theory is *true*. For others, it could for example refer to uncertainty about the moral view one would adopt after more reflection.

[2] More precisely, MacAskill (2016) proposes a version of Borda's rule modified in a way that behaves better in the presence of clones.

which actions are permissible, but give no further information; other theories may mainly be interested in forbidding certain actions, for example because they violate rights. Thus, for these settings, we need aggregation mechanisms that are able to combine inputs that come in fundamentally different formats.

Some voting rules may admit natural generalisations that allow this combination of different input formats. Range voting comes to mind: this rule would 'cardinalise' any ordinal input in some natural way, and sum the results. But such rules lack an axiomatic study, and it is not clear how exactly the cardinalisation should be performed.

Some people with uncertainty over moral theories may have all or most of their credences in moral theories that have a cardinal structure, such as utilitarian proposals. Such decision-makers need an aggregation rule for cardinal preferences. However, most of voting theory has considered the ordinal case only to avoid having to make inter-personal comparisons of utility, and pointing to concerns that it might be cognitively difficult for voters to figure out their own cardinal preferences, as well as concerns about avoiding strategic behaviour. The latter concerns are much less applicable in the setting of moral uncertainty. Since decision-makers could benefit from having cardinal aggregation procedures for this setting (and others), more formal analysis is needed.[3]

When voting over which restaurant to go to, say, it is not uncommon to observe every possible preference ranking to be reported by some agent. In our case of moral theories, we should instead expect the input to be *structured*: most logically possible moral rankings are implausible or inconsistent and should receive (almost) zero credence. Maybe this structure can be used to evade impossibility results and design better aggregation rules, like is possible in the case of single-peaked preferences (Black, 1948). For example, this approach may be relevant to the problem of aggregating different *population axiologies*, first considered by Greaves and Ord (2017). A population axiology ranks states of the world with different numbers of people and different welfare levels in order of goodness. Classical examples are *total* and *average utilitarianism*, which hold that a state is better than another iff the sum (resp. average) of the individual welfare levels is higher in the first state. There is a well-developed theory of properties that a single axiology should satisfy (Blackorby et al., 2005), and these properties define domain restrictions for the problem of aggregating several axiologies. To analyse this setting, results from the literature on social choice on economic domains could be applicable (see Le Breton and Weymark, 2010).

Further Problems. Many other useful contributions of economic design to the project of Effective Altruism are imaginable. Toby Ord's (2015) article on *Moral Trade* is particularly inspirational in this regard. One problem Ord considers is that of zero-sum conflicts in charitable giving: for example, in political campaigns, if person 1 gives $1 to party A and 2 gives $1 to opposing party B, then these donations may 'cancel out'; both might prefer that the total $2 would go to a neutral charity instead. One can imagine a

[3] For some axiomatic results, see Pivato (2014) for a characterisation of range voting and 'formal utilitarianism' as the 'most-expressive' voting rules satisfying reinforcement, as well as some work on relative utilitarianism (Dhillon, 1998; Dhillon and Mertens, 1999; Börgers and Choo, 2017).

system which would match such donors to avoid the zero-sum contributions, but there are several problems such a system would need to overcome, including strategic manipulation and trust issues (Ord, 2015). Christiano (2016) gives an example of a mechanism that could work better than the naive one, but it requires the market maker to contribute additional money. A formal analysis of this setting could yield interesting insights.

Many EAs want to choose a career that maximises their altruistic impact. Often, people try to find their *comparative advantage* in choosing a certain job. But it is difficult to know what one's comparative advantage is without having a lot of knowledge about other people's abilities. To help with this, it would be interesting to study matching mechanisms that have very low communication complexity: they only need to query few pairs of participants and jobs for their suitability, in order to determine a job assignment that is approximately optimal.

If a group of people has access to a shared donation pot (e.g., the employees of a company), how should they decide where to donate this money as a group? Voting seems like the right tool to make this decision; if the donation pot can be split, *probabilistic* voting rules (see Brandt, 2017) could suggest a good outcome. Which rules can we recommend for this situation?

Conclusions. In this article, I have outlined several ways in which theorists in the area of economic design can support the Effective Altruism community. Progress on these problems may have high impact and improve many lives. In addition, the problems are technically and conceptually interesting, and may have applications in other areas. I am excited to see what we can do.

References

D. Black. On the rationale of group decision-making. *The Journal of Political Economy*, 56(1):23–34, 1948.

C. Blackorby, W. Bossert, and D. J. Donaldson. *Population issues in social choice theory, welfare economics, and ethics.* Cambridge University Press, 2005.

T. Börgers and Y.-M. Choo. A counterexample to Dhillon (1998). *Social Choice and Welfare*, 48(4):837–843, 2017.

F. Brandt. Rolling the dice: Recent results in probabilistic social choice. In U. Endriss, editor, *Trends in Computational Social Choice*, chapter 1. AI Access, 2017. to appear.

Centre for Effective Altruism. Fundraising Prospectus – Winter 2017, December 2016. https://www.centreforeffectivealtruism.org/fundraising/.

P. F. Christiano. Repledge++, October 2016. https://sideways-view.com/2016/10/31/repledge/.

V. Conitzer and T. Sandholm. Expressive markets for donating to charities. *Artificial Intelligence*, 175(7-8):1251–1271, 2011.

A. Dhillon. Extended Pareto rules and relative utilitarianism. *Social Choice and Welfare*, 15(4):521–542, 1998.

A. Dhillon and J.-F. Mertens. Relative utilitarianism. *Econometrica*, 67(3):471–498, 1999.

E. Elkind and A. Slinko. Rationalizations of voting rules. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia, editors, *Handbook of Computational Social Choice*, chapter 8. Cambridge University Press, 2016.

GiveWell. Against Malaria Foundation – November 2016 version, November 2016. https://www.givewell.org/charities/against-malaria-foundation/november-2016-version.

H. Greaves and T. Ord. Moral uncertainty about population ethics. *Journal of Ethics and Social Philosophy*, 2017. to appear.

H. Karnofsky. The value of coordination. The GiveWell Blog, December 2014. https://blog.givewell.org/2014/12/02/donor-coordination-and-the-givers-dilemma/.

M. Le Breton and J. A. Weymark. Arrovian social choice theory on economic domains. In K. J. Arrow, A. K. Sen, and K. Suzumura, editors, *Handbook of Social Choice and Welfare*, volume 2, chapter 17. Elsevier, 2010.

W. MacAskill. Normative uncertainty as a voting problem. *Mind*, 125(500):967–1004, 2016.

W. MacAskill. Effective Altruism: Introduction. *Essays in Philosophy*, 18(1):1, 2017.

H. Moulin. Condorcet's principle implies the no show paradox. *Journal of Economic Theory*, 45(1):53–64, 1988.

L. Muehlhauser. Technical and philosophical questions that might affect our grantmaking. Open Philanthropy Project, March 2017. https://www.openphilanthropy.org/blog/technical-and-philosophical-questions-might-affect-our-grantmaking.

T. Ord. Moral trade. *Ethics*, 126(1):118–138, 2015.

M. Pivato. Formal utilitarianism and range voting. *Mathematical Social Sciences*, 67: 50–56, 2014.

A. E. Roth, T. Sönmez, and M. U. Ünver. Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488, 2004.

B. Todd. The value of coordination. 80,000 Hours, February 2016. https://80000hours.org/2016/02/the-value-of-coordination/.

H. P. Young. Condorcet's theory of voting. *American Political Science Review*, 82(4): 1231–1244, 1988.