Journal of Institutional Economics (2024), **20**, e41, 1–18
doi:10.1017/S1744137424000146

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

The paradox of effective altruism

Tomasz Kopczewski[1] and Iana Okhrimenko[2]

[1]Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland and [2]Department of Econometrics, Lazarski University in Warsaw, Warsaw, Poland
Corresponding author: Tomasz Kopczewski; Email: tkopczewski@wne.uw.edu.pl

(Received 1 June 2023; revised 7 May 2024; accepted 8 May 2024)

Abstract

Normative reasoning within the mainstream economic framework has been largely shaped by utilitarian ethics. The growing popularity of effective altruism indicates that the utilitarian spirit has also permeated the sphere of social sentiment, evaluating our pro-social behaviour and charitable giving in terms of efficiency. The present study assesses the appropriateness of judging social outcomes through the prism of allocative efficiency by questioning to what extent the society of effective altruists is robust, sustainable, and resilient. Using computer simulations based on the dictator game, we demonstrate that a society of welfare-maximising effective altruists can achieve an optimal outcome alongside equality under extremely restrictive assumptions, such as the uniformity of giving strategies (i.e. interacting with other effective altruists exclusively) and initial equality of wealth distribution. Yet, in the world of unequal opportunities, utilitarian giving tends to increase wealth disparity. In addition, in polymorphic societies, effective altruists underperform compared to deontological (or unconditional) altruists. Consequently, we demonstrate that striving for allocative efficiency might undermine the equality and resilience objectives and question whether the former should remain the dominant economic normative principle.

Keywords: agent-based modelling; effective altruism; evolutionary economics; institutions; social heuristics

Introduction

In mainstream economics, most normative attributions are made on the basis of utilitarian ethics. The utilitarian idea of 'maximising the aggregate good' can be found in public policy (Goodin, 1995, 1998; Little, 2002; Riley, 2008) and public choice theory (Arrow, 1951; Hillinger, 2004). The effective altruism movement advocating rational and goal-directed giving is becoming increasingly popular among scholars (Genç *et al.*, 2020; MacAskill, 2018, 2022; Singer, 1972, 2009). Besides, it is gaining followers outside the academia: Bankman-Fried (formerly one of the wealthiest crypto executives), Elon Musk (founder of Tesla, Neuralink, and The Boring Company), and Dustin Moskovitz (the co-founder of Facebook) are among the most active advocates of effective altruism in media. It seems that the utilitarian spirit has also permeated the sphere of social sentiment, evaluating our pro-social behaviour and charitable giving in terms of efficiency.

Advocates of effective altruism claim that rationally targeted charitable giving (i.e. the pattern of altruism that seeks to achieve the greatest good with limited resources) will lead to the greatest gains in aggregate welfare. They also claim that targeted charitable giving can ameliorate the significant inequalities in levels of well-being. We wonder whether it is necessarily true that such targeted giving will have the wealth-maximising and egalitarian effect relative to less targeted forms of charity, such as deontological giving. Our results suggest that in some very rare and limited circumstances, effective altruism performs better on these dimensions. However, paradoxically, under more realistic

assumptions, deontological altruism tends to perform better in terms of wealth maximisation and egalitarian outcomes.

We use a computer simulation as our primary research tool. The simulation is similar to Andreoni and Miller's (2002) *Giving according to GARP* (the general axiom of revealed preferences) laboratory experiment[1]: agents allocate scarce resources between themselves and others under varying budget constraints, except that in our version, endowment, pass, and hold values are randomly generated each round. Agents decide how much to give according to their 'ideal type' (i.e. deontological, effective, and zero-intelligent altruists). The outcome of the agents' interaction (the value of the endowment and the equality of its distribution) in monomorphic and (more critically) polymorphic societies serves as a basis for inferring the relative performance of the different giving strategies.

The contribution of this paper is twofold. First, it attempts to demonstrate the shortcomings of the teleological orientation and the emphasis on allocative efficiency of effective altruism using a practical tool (in contrast to a substantial body of literature based on purely theoretical arguments). We also propose an alternative criterion for evaluating social outcomes, namely resilience. Furthermore, to our knowledge, there have been limited attempts to conceptualise the principles of utilitarian and deontological altruism as 'giving strategies' – we strive to fill the knowledge gap.

Utilitarian ethics, effective altruism, and efficient distribution of resources

Teleological ethics refers to the system of moral reasoning that takes the consequences of an action as the primary basis for its judgement. Utilitarianism, which asserts utility maximisation as the greatest good, is one of the most vivid teleological ideas. Although it is not easy to trace the history of the utilitarian concept to its roots, Bentham (1834a, 1834b) was perhaps the first modern scholar to formulate the principles of utilitarian ethics explicitly. According to Bentham, the pursuit of the greatest happiness is central to human motivation, and this fact is sufficient to establish the principle of 'goodness as utility'. Empathy, compassion, and social instincts are explicitly interpreted as acts of individual self-optimisation: 'But the pleasure I take in the prospect of giving pleasure to my friend, whose pleasure is it but mine?' (Bentham, 1834a: 148). This quote perfectly describes how most modern economists evaluate altruism and pro-social behaviour using the notion of social (or other-regarding) preferences.

Mill's (1863) contribution to utilitarianism is to refine and extend the principle of greatest happiness proposed by Bentham. Mill introduces the concept of the quality of pleasure and posits that individuals who have the capacity to experience both mental and physical pleasure will tend to prefer the former. Bentham (1834a) argues for the justification of utilitarian motivation based on the belief that it is inherent in human nature, whereas Mill (1863) advocates utilitarianism as a widely accepted moral principle because of its beneficial effect on the collective outcome.

Utilitarian ethics has significantly shaped the development of normative attributions in mainstream economics (Colander, 2000). The maximisation of social happiness (sometimes described as the efficient use of scarce resources in the context of collective choice) is implicitly assigned a role as the basis for any normative judgement, although economics is claimed to be value-neutral (see Arrow, 1997; Robbins, 1938). Of course, it would be a mistake to say that economics does not deal with various considerations that seem to have nothing in common with narrowly defined self-interest; on the contrary, they play a prominent role being conceptualised as economic goods. The notions of 'utility' and 'social welfare' are extremely complex domains that encompass a wide range of economic goods, such

[1]Rational preferences are consistent across different budget constraints; this consistency can be tested on the basis of the general axiom of revealed preferences (GARP), the weak axiom of revealed preferences (WARP), or the strong axiom of revealed preferences (SARP). Treating altruism as an economic good implies that altruistic preferences can be tested for rationality (understood as consistency of choices). Andreoni and Miller (2002) distributed some tokens (endowment) among participants. They could keep or give them away under different trade-off values (e.g. a token could be worth two income units for oneself and three income units for others), which can be understood as the price of altruism. Therefore, altruistic preferences could be tested for consistency using the GARP, WARP, or SARP (see detailed explanation in Appendix A).

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

as altruism (the 'warm glow of giving' is something for which we sacrifice our own resources), leisure, justice, equality, and so on. Consequently, when these considerations appear in economic analysis, they are only important as long as they are valued (i.e. as long as they contribute to individuals' utility).

It is worth mentioning that utilitarianism is typically juxtaposed to egalitarian motives. Rawls (1971) pointed out that aggregate welfare maximisation does not necessarily benefit the most deprived individuals, thus contradicting the idea of fairness defended by Rawls from the 'original position' standpoint. Sen (1972) points out that although utilitarianism gained a reputation as an egalitarian movement (perhaps because assuming identical preferences and diminishing marginal utility, distribution according to the marginal benefit means assigning more resources to the least wealthy individuals – see Brandt, 1979, for more details) its innate nature has nothing to do with equity (or equality) concerns. Stark *et al.* (2012) point out that if individuals care about their position in the income distribution, utilitarian and egalitarian social planners will arrive at the same end state (indeed, assuming that all individuals are characterised by strand inequality aversion, the goals of utility maximisation and equality promotion will coincide). Galanis and Veneziani (2022) demonstrate that distribution implemented by the benevolent utilitarian planner decreases income inequality over time, provided that individuals exhibit reference-dependent preferences – yet, the results are, to a great extent, constrained by the diminishing marginal utility of wealth assumption, thus not immune to Sen's (1972) critique.

Today, we are witnessing a new round in the evolution of the act-utilitarian spirit – effective altruism. Effective altruism implies that a society can be better off by incrementally redistributing an endowment. As Singer (1972) points out, the sacrifice of a luxury (e.g. lipstick) does not have the same moral significance as human suffering due to the lack of essential goods (such as food, medicine, and shelter). Therefore, whenever someone faces a choice between buying a new lipstick and donating to charity, charity is a moral obligation. Singer's (1972) argument implies a redistribution of endowments based on the principle of maximum marginal utility: whoever enjoys the most should get the most (obvious for any economist, such a pattern of altruistic giving is not only *effective* but also *efficient* – see more details in Appendix A).

The above premise is easy to challenge from the perspective of economic epistemology. Such a transaction requires an analysis of the marginal utilities of different agents associated with different goods or an interpersonal utility comparison (IUC). From the collective choice theory standpoint, IUC could imply that one's own preferences should be given more weight than those of others (Arrow, 1951; List, 2003). Moreover, it is impossible to experimentally verify the relative values of the marginal utilities that different individuals associate with various goods (Myerson, 1983). IUC is not a valid concept for normative economics. Nevertheless, introducing IUC is the only way to allow welfare economics to move beyond trivial Pareto improvements, which is why most welfare economists do so (Hammond, 1991; Johansson-Stenman, 1998). On the other hand, Harsanyi (1978) is in favour of the 'general possibility' of the IUC on the grounds that under certain conditions (such as cardinal preferences and impartiality[2]) it can be meaningful. Moreover, provided that individuals are rational in the Bayesian sense, the social welfare function takes the form of an arithmetic mean of individual utilities, which justifies utilitarian morality from the standpoint of rationality.

The reliance on the IUC in Singer's (1972, 2009) normative attribution model can be justified by incorporating the objective criterion (or self-sufficient truth) as defined by Scanlon (1975). A meal will provide more benefit to a hungry person than a new lipstick will to anyone, and this statement can be qualified as a self-sufficient truth. The problem is that most effective altruists apply the same principle to much more complex situations, implying that an objectively better option can still be identified. Distributing malaria nets seems a brilliant, cost-efficient way of fighting malaria – until one figures out they are actually used for fishing, contributing to environmental degradation in Africa (Jones and Unsworth, 2020). MacAskill (2018) acknowledges that such 'epistemological issues' remain an

[2]According to Harsanyi, impartiality refers to making a decision about a social outcome without knowing one's own social position after the decision has been made (see also Harsanyi, 1955). This procedure is known as the 'veil of ignorance', although this name was popularised by Rawls (1971) and not originally used by Harsanyi.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

open question to be resolved at some point in the future, which resembles Bentham's (1823) felicity calculus exercises.[3] The impossibility of assessing subjective utility or predicting the remote consequences of an action can hardly be treated as secondary concerns that will inevitably be addressed at some point; rather, they represent the major flaws in utilitarian thinking.

For the most modern version of effective altruism, teleological orientation – the additional key component of utilitarian ethics – takes a rather extreme form. MacAskill (2022) advocates longtermism, proclaiming that it is our ultimate moral obligation to care for the welfare of all future generations. Although it is hard to argue with the above statement, being a true longtermist would require something more: a clear vision of the desirable 'end state', a belief that such a state is either favoured by the rest of society (or can be undoubtfully identified as objectively superior to all alternatives), and (last but not least) living in a world free of radical uncertainty.[4] Overall, longtermism mirrors a well-known 'engineering' approach to social organisation (Hayek, 1944).

It would be too ambitious to argue that there is no boundary between effective altruism and classical utilitarianism. The former is claimed to be more nuanced and complex (MacAskill, 2018). However, the definition of effective altruism proposed by MacAskill (2018) highlights maximising, science-alignment, tentative welfarism, and impartiality, which should sound recognisable to anyone familiar with collective choice problems being discussed under the neoclassical economics shield. In addition, most of the hypothetical cases described by advocates of effective altruism typically resemble the familiar problem of redistribution based on the marginal utility criterion (Elder and Fischer, 2017; Gabriel, 2016; Singer, 1972, 2009). The treatment of efficiency as the main normative criterion and the teleological orientation are innate to effective altruism and remain the key focus of this paper.

Deontological altruism and long-term resilience criterion

Deontological ethics is juxtaposed with teleological moral thinking in the sense that the former disregards the consequences of the act as the basis for its evaluation. Unlike utilitarianism (often described as a product of the French Enlightenment – see Hayek, 1945), deontology (as we understand it nowadays) can be traced back to ancient Greek philosophy (see Russell, 1946).[5] Among modern philosophers, Kant (1785) is, perhaps, the most famous advocate of deontological ethics, who objects to consequentialist moral reasoning[6] and believes that all humans have specified moral duties, which they are obliged to follow because there is nothing to justify the violation of these duties.

Unlike effective altruism, altruism rooted in deontological reasoning might be tricky to define. Van Staveren (2001) asserts that 'deontological ethics is about following universal norms that prescribe what people ought to do, how they should behave, and what is right or wrong' (p. 23), aligning with Kant's 'rule worship'. McNaughton and Rawling (2007) postulate that 'deontologists characteristically hold that we must not harm people in various ways', although 'deontology sometimes requires agents not to maximise the good'; the former premise sounds in line with deontological (although not Kantian) logic, whereas the latter one resembles teleological thinking. According to White (2009), deontology implies that 'not all ethical judgements regarding actions can be made on the basis of outcomes or consequences' (p. 300), assuming that in some circumstances, teleological reasoning can be

[3]Bentham claimed that the net benefit of the action could be given a precise value based on seven criteria (intensity, duration, certainty/uncertainty, remoteness, fecundity, purity, and extent). The practical application of such a framework is questionable.

[4]We refer Keynes's (1921) notion of radical uncertainty. Therefore, the absence of radical uncertainty implies that agents are aware about all the possible outcomes and are able to detect their probabilities: there are no 'black swans' on the horizon.

[5]The reader might argue that utilitarianism has much in common with Epicurus's hedonism, because both regard pleasure as the greatest good and recognise the superiority of higher pleasures (such as intellectual satisfaction, the feeling of companionship, etc.). However, there is a significant difference. Hedonism concentrates on private happiness, whereas utilitarianism is concerned with society as a whole, with the aggregate good. Indeed, the latter suits the 'engineering approach' arising from Cartesian logic (and criticised by Hayek, 1945).

[6]'A good will is good not because of what it performs or effects, not by its aptness for the attainment of some proposed end, but simply by virtue of the volition […]' (Kant, 1785).

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

justified. Sunstein (2014) affirms that 'deontologists believe that some actions are wrong even if they have good consequences'. Hardt (2020) argues that deontological norms are not absolute, and sometimes, violation of perfect duties can be justified. Our intuition is that 'pure' deontological reasoning (in the Kantian sense) requires a belief in fundamental moral rules; being religious is not crucial, but it certainly helps. Consequently, modern views of deontology include a teleological component, in the sense that moral duties should be consistent with general welfare goals to be justifiable.

To depict deontological altruism in simulations, we need to understand how individuals manifest it in social exchange. Sunstein (2014) describes deontological reasoning as a kind of 'moral heuristics': in situations of uncertainty and time pressure, we rely on 'rapid, automatic, emotional processing', following specific unconditional rules that take the form of formal (e.g. legally enforced) and informal (i.e. social norms[7]) institutions. Thus, in the present context, deontological altruism is seen as the exercise of social norms by individuals, while the particular social interaction strategy represents a social norm as such. A big question is the extent to which social norms are actually unconditional; for example, Sugden (2018) highlights that our tendency to follow social norms depends on whether others are observing us. However, even though social norms may be constrained by external approval or disapproval, they have the deontological component of acting in a certain way regardless of the expected consequences.

The deontological (heuristical) giving has been known since ancient times. Tithing is mentioned in the Old Testament; Judaism requires at least 10% of income to be given to charity (this practice is common among Muslims and Christians). Singer (2009) refers to this practice and argues that one can give a tenth of one's income to effective charitable foundations. Yet, ancient religious communities had little in common with modern charitable foundations that cherish the idea of effective altruism. The former served as orphanages, provided schooling and helped the poor by collecting and distributing donations. However, it would be more accurate to think of them as charities of 'spontaneous order' (in Hayek's sense) rather than the product of deliberate design with a clear end goal. Religious confessions provide communities with public goods and help to ensure organised collective action; they stay resilient because they do not have a central planner striving for the 'optimal design' of charity and other forms of human interaction (Gill, 2020).

Human beings are social creatures; as long as they can feel the ties with their community and a proper sense of sociability, altruism can be effectively institutionalised (Kropotkin, 1904). Thus, although Singer (2009) attempts to fit tithing custom into the paradigm of effective altruism, there is no more profound connection between the two. Although we do not feel sufficiently confident to provide an explicit definition of deontological altruism, intuition dictates that it can be described as sacrificing a share of one's income for the sake of giving without concern for how efficiently the donation is used.

Our take on conceptualising ethical systems in action allows for drawing a borderline between utilitarian and deontological altruism. However, the distinction between deontology and virtue ethics[8] is much less pronounced. Intuitively, there is a difference between following the 'logic of duty' and developing a virtuous personality through acting accordingly. Yet, the aforementioned difference refers to motivation, not action. The practice of tithing might arise from one's devotion to duty (expressed as a religious norm or social convention) or the desire to be a good person. Therefore, although virtue ethics represents a separate ethical system, it mostly remains beyond the scope of our interest.

Overall, we concede that deontological giving might lead to inefficient allocation of scarce resources, but argue that it might serve as a better giving pattern from another perspective, namely resilience and sustainability criteria. Gill and Thomas (2023) also doubt the appropriateness of the normative criterion of allocative efficiency. Gift-giving inevitably leads to deadweight losses and

[7]Informal institutions can be defined as uncodified rules governing key aspects of social interaction. Studies of social exchange come from a variety of fields, which makes it difficult to use consistent terminology. Overall, however, 'social norms', 'social heuristics', 'cultural norms', and 'informal institutions' are close enough to be used interchangeably (see, for instance, Alesina and Giuliano, 2015).

[8]Aristotle's (2014) *Nicomachean Ethics* represents the oldest and the most famous work in defence of virtue ethics: 'men become gods by excess of virtue, of this kind must evidently be the state opposed to the brutish state' (Book IV).

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

represents, therefore, an undesirable practice from a mainstream economic perspective; cash transfers would be much more efficient. However, gift-giving makes sense once one abandons the mainstream economic view. Such custom can be seen as a symbolic sacrifice (sometimes involving the deliberate destruction of value) caused to develop mutual trust (primarily by demonstrating one's trustworthiness).

Giving gifts violates allocative efficiency but creates and maintains an environment essential for further cooperation (including market transactions). One would hardly believe that gift-giving is motivated by the desire to sustain the important social practice and thus contribute to social well-being in the long-run perspective. Instead, a giver follows a common practice – some heuristics – developed in the process of social evolutionary selection. Perhaps one might say that being an effective altruist requires the belief in the superiority of deliberate design, while deontological ethics implies following the rule, not necessarily realising its greatest purpose (or even the existence of such).

Simulation design

The simulation mirrors the *Giving according to GARP* (Andreoni and Miller, 2002) experiment. The original experiment includes eight rounds; the participants are randomly assigned different partners during each round. All participants receive the same endowment (the number of tokens) to allocate between themselves and others under various 'hold' and 'pass' values. Under such a setting, participants face a trade-off between two economic goods: own pay-off and altruistic giving under various budget constraints. Consequently, the original study concentrates on testing social preferences for consistency with the GARP (see Appendix A).

The present study adopts a similar experimental design for a different purpose: evaluating the various giving strategies (including effective altruism and deontological altruism) based on multiple criteria. From the narrow choice-consistency perspective, deontological altruism violates the principles of rational choice (in contrast to effective altruism – see Appendix A). Besides, intuition dictates that a society of effective altruists should have higher aggregate wealth than the deontological altruists (results supporting this intuition are presented in further sections). From the mainstream perspective, effective altruism seems the most desirable form of social exchange from society's standpoint. Yet, it is worth examining the problem from a different perspective by observing the outcomes of interaction between agents following different giving strategies in various circumstances.

Our simulation was inspired by so-called 'battle of algorithms' studies exploring the relative performance of various decision-making rules (Axelrod, 1987; Imhof *et al.*, 2007; Lomborg, 1996; Nowak and Sigmund, 1993; Wahl and Nowak, 1999) in which pseudo-intelligent agents interact according to fixed rules constrained by their type. In the present study, agents' types are defined by their giving strategies; the decision-making framework (i.e. endowment alongside the hold and pass values) changes every round.

Each iteration of the simulation contains two loops. In the outer loop, endowment (in the form of tokens) is drawn from the range [10, 100] for a population of N for R periods (rounds), generating random matrix $M[N, R]$. Therefore, wealth distribution is not uniform at the beginning of the game. We specify eight rounds (analogously to the original experiment), although the number of rounds can be amended without any noticeable change in the results. In each round, each agent interacts with a new partner. The monomorphic populations consist of 1,000 agents; each additional type added to the society increases the population by 1,000 agents. In polymorphic populations, all the types are combined; because the partners are selected randomly, the agents can interact with representatives of various types (including their own). During the game, the population composition remains unchanged. Then, the hold (V_s[N, R] – 'payment to self') and pass (V_o[N, R] – 'payment to others') values are randomly picked from the range [1, 10]; these values are uniform for all the agents, regardless of their type.

It should be noted that although, technically, one is free to specify any range for the pass and hold values, the behaviour of effective altruists is sensitive to the ratio of the aforementioned, because it

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

constitutes parameter f (see the detailed discussion in the next section). Therefore, f should be specified considering the maximum gap between the pass and the hold value. Otherwise, assuming that the parameter value is too high, effective altruists might donate nothing (and vice versa).

The inner loop generates agents' choices (i.e. the number of tokens to pass) under the budget constraints specified by rows of matrixes $M[N, R]$, $V_s[\cdot, R]$, $V_o[\cdot, R]$ depending on the given agent type (the behaviour of each type is described in the next section). We use Monte Carlo simulations for the sake of robustness. In each of the 100 iterations, information about the total wealth of individuals and the entire society is collected, and the Gini coefficient is used to measure wealth inequality.

Giving strategies

The game design utilised in the simulation serves as a simplified model of social exchange. Therefore, to conceptualise effective and deontological altruism in the simulation, one should take the aforementioned ideas to the extreme and concentrate on their most distinct aspects. Teleological orientation and allocative efficiency are distinguishable features of effective altruism. In *Giving according to GARP* setting, such a decision-making principle can be only conceptualised through giving according to the marginal benefit rule (although in a more complex real-world environment, effectively altruistic charity might not even concern the proportion of income being donated – Côté and Steuwer, 2023). Consequently, this paper does not attempt to criticise effective altruism in a holistic manner; instead, we address the implications of teleological reasoning and treating efficiency as the principal normative criterion exclusively.

Similarly, when referring to deontological altruism in simulation, we emphasise its unconditional and heuristic nature – a deliberate simplification of a more complex concept for simulation purposes. It should be stated clearly that we see deontological altruism as a spontaneously developed and constantly evolving phenomenon. Our natural inclination to give and help is cultivated through socialisation,[9] but this does not mean that we cannot learn or change a giving strategy depending on external environment constraints. At the macro level, mutual expectations and shared social norms may change[10]; at the micro level, individuals adjust their interaction strategies according to the signals they receive. As we are unable to include the above aspects in the simulation, we should nevertheless warn that our deontological altruists are quite dumb compared to their hypothetical and more realistic version.

In the simplified simulation setting, we distinguish between three types of agents, namely:

(1) *Type 1: deontological altruists*. Deontologically altruistic agents always pass the fixed portions of their token endowment, regardless of the hold and pass value. They pass $t \times M_R$ and hold $(1 - t) \times M_R$ tokens in each round R. Two extreme attitudes can be distinguished: unconditionally selfish agents tend to pass nothing ($t \to 0$), whereas pathological altruists[11] tend to pass the entire endowment to other players ($t \to 1$).
(2) *Type 2: effective altruists*. Welfare-maximising effective altruists distribute their token endowment in line with the allocative efficiency principle. When the hold value exceeds the pass value, they hold all the tokens, and vice versa: when the pass value exceeds the hold value, they pass all the tokens. However, other sub-types might be more or less sensitive to efficiency

[9]We are well aware of arguments in favour of the 'innate' nature of altruistic motives (see, for instance, Haidt, 2012). However, such a distinction makes no major difference in the present context, because 'innate' modules are also believed to be subject to evolutionary selection.

[10]Social heuristics (or informal institutions) are not supposed to be 'good' or 'optimal'; rather, they facilitate survival in the complex external environment. For example, a low level of trust reduces cooperation and, consequently, trade and exchange. And yet, in a totalitarian society, the strategy of trusting no one is the one that increases the likelihood of survival, so, it becomes common.

[11]We adopted the term 'pathological altruism' from Oakley *et al.* (2011): 'in essence, pathological altruism might be thought of as any behaviour or personal tendency in which either the stated aim or the implied motivation is to promote the welfare of another. But instead of overall beneficial outcomes, this altruism instead has irrational (from the point of view of an outside observer) and substantial negative consequences to the other or even to the self'.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

considerations. The parameter f stands for a psychological multiplier indicating how many times the pass value must be greater than the hold value so that the agent is willing to give (we refer to it as the 'efficiency considerations parameter'). The condition for transferring endowment is $f \leq V_o/V_s$. Three essential strategies can be distinguished. Agents for whom parameter f approaches infinity ($f \to \infty$) exhibit egoistic tendencies: they donate if the pass value is substantially higher than the hold value. Agents with f approaching zero ($f \to 0$) feature pathological altruism because they only hold tokens if the hold value surpasses the pass value enormously. Welfare-maximising effective altruists operate on a trade-off principle: they are concerned with aggregate welfare maximisation only, implying that $f = 1$.

(3) *Type 3: zero-intelligence altruists.* Zero-intelligence agents assign a random number of tokens to other players regardless of the hold and pass value trade-off, constrained solely by their token endowment (analogously to Sunder's, 1993, agents).

Interaction in monomorphic societies

In monomorphic societies, the agents interact with their type solely. Each society consists of 1,000 agents (this parameter was selected arbitrarily and can be amended without any implications for the results). The particular pattern of giving is evaluated based on the value of the endowment and its distribution; we utilise the cumulative distribution functions (CDFs) for this purpose (see Figure 1). The interpretation is as follows: for any value of endowment (x-axis variable), the function represents the proportion of the population (y-axis variable) with endowment below the aforementioned value. It is possible to say that one type stochastically dominates the other if, for any cumulative proportion of the population, such a type possesses the greater endowment (in this case, the endowment distribution of the dominant type is located to the right-hand side of the distribution of the dominated type). The results of the simulations in the form of wealth distributions featured by each 'ideal type' and society as a whole are constrained by the values of the parameters t and f. By manipulating these variables, one can create infinite options, and we cannot discuss (or even demonstrate) all of them in the present paper. Therefore, we present the selected results – the most interesting and illustrative ones in our view. However, one could check the results in monomorphic and polymorphic societies, assuming different values of the key parameters using the online appendix.[12]

The outcomes in the societies of zero-intelligence altruists (who select the number of tokens to be passed randomly) are treated as a benchmark for assessing the remaining types. In the first and the second panels, we design relatively selfish deontological and effective altruists (assuming that the former donate a relatively small proportion of their endowment and the latter require a substantially greater value of pass value compared to the hold value to donate their entire endowment to others). In the simulation displayed in panel 3 (corresponding to $f = 1$ and $t = 0.5$), effective and deontological altruists are designed to be indifferent between their own and others' pay-off (the total pay-off maximisation objective guides the former type, and the latter type passes half of the endowment). Panel 4 represents the scenario of pathological altruism among the effective and deontological agents.

Panel 1 shows that radically selfish deontological and effective altruistic giving produce similar results to zero-intelligence agents. However, effective altruists stochastically dominate zero-intelligence and deontological altruists in all other cases. In contrast, about half of the deontological altruists underperform even zero-intelligence agents in all but the first simulation. Interestingly (yet not surprisingly), welfare-maximising effective altruists (i.e. featuring $f = 1$) accumulate the greatest wealth compared to the effective altruists featuring higher or lower values of the 'efficiency considerations parameter'.

Effective altruists dominate deontological altruists with similar levels of generosity in isolation. At the same time, however, the latter pattern of giving appears to be the most egalitarian. The distribution of changes in Gini coefficients in 100 iterations (see Appendix B, Figure B1) shows that the

[12] https://microeconomics.shinyapps.io/altruism_simulation/#section-intro.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

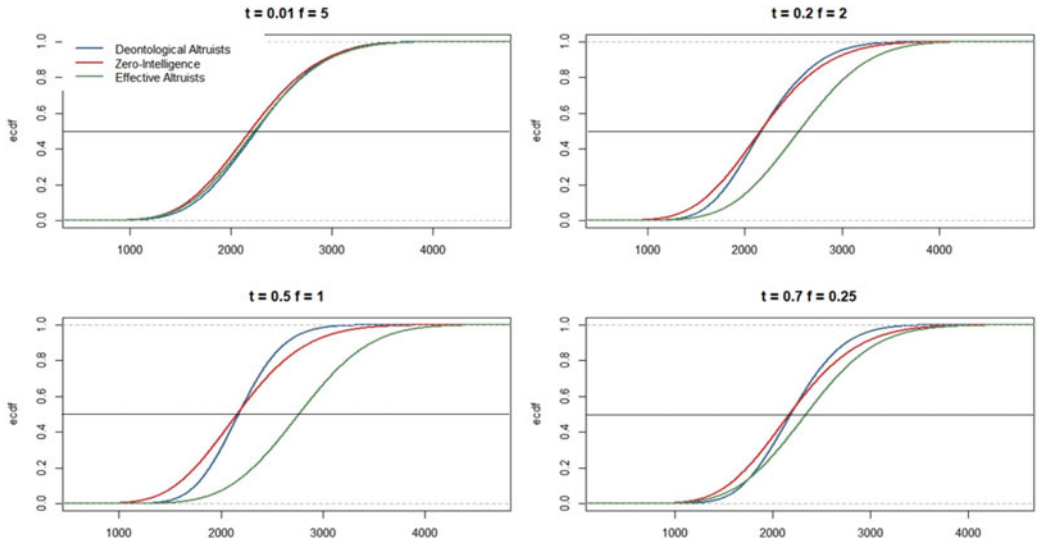

Figure 1. Wealth distributions in monomorphic societies of deontological altruists, effective altruists, and zero-intelligence altruists assuming different values of t and f parameters.

Note: In every panel, the x-axis represents the simulated wealth of society after the exchange, and the y-axis shows the proportion or percentage of data points that are less than or equal to the corresponding wealth values. The point where the curve crosses 0.5 on the y-axis (indicated by the black line) corresponds to the median. If the curve rises quickly, much of the wealth is clustered around lower values. A slower rise indicates wealth is more evenly distributed across the range. The 'fat' curve's tails indicate the unequal distribution and outliers. Comparing distributions of simulated societies: when lines diverge, it indicates differences in the distribution. The further to the right the line is, the more wealth a given society has. The parallel shift of a given distribution (A) about another distribution (B) to the right indicates first-order stochastic dominance: for every possible income level, the proportion of the population with income below that level is smaller in (A) than in (B).

deontological pattern of giving reduces the degree of initial wealth inequality. In contrast, effective altruists experience increased inequality (although not as substantial as in the population of zero-intelligence agents).

It is worth mentioning that in one of the preliminary versions of the simulation, we assumed the uniform distribution of initial endowment. In such a setting, the isolated population of welfare-maximising effective altruists featured the lowest wealth inequality among all the groups *ex-post*. Therefore, in the ideal world of perfect equality of opportunity (captured by equality of wealth), where all individuals are concerned only with the common good (maximising aggregate wealth), equality and wealth are not competing but complementary concerns. Such a society would indeed be a true utopia. However, in a more realistic setting, where the distribution of endowments is uneven, effectively altruistic giving leads to increasing wealth inequality. Although this is not the primary objective of the paper, our robust simulation results show that the goals of total wealth maximisation and equality can be aligned in rare circumstances only (e.g. under the assumptions of Stark *et al.*, 2012, or Galanis and Veneziani, 2022). In general, however, the utilitarian distribution leads to the anti-egalitarian outcome.

Interaction in polymorphic societies

The polymorphic societies consist of deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions. The populations include 1,000 agents of each type; analogously to the monomorphic groups' analysis, the results are robust to the changes in the population size parameter N. The simulations presented in Figure 2 were designed according to the same logic as the simulations reported in Figure 1: our goal was to compare the performance of the deontological and

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

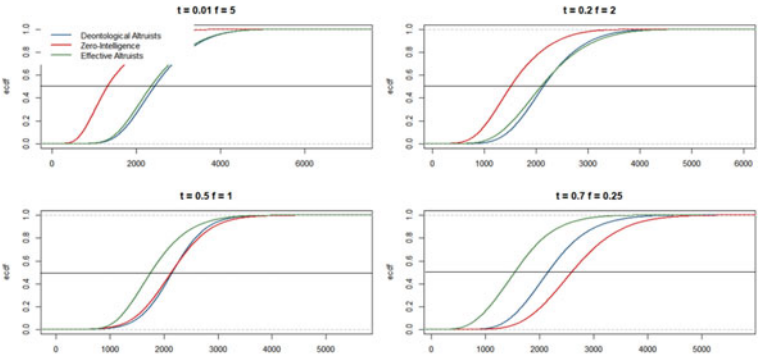

Figure 2. Wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming different values of t and f parameters.

effective altruists featuring a comparable degree of generosity. Allowing agents to interact with other types caused radically different outcomes. Although the shape of CDFs of relatively selfish effective and deontological altruists did not change significantly, zero-intelligence altruists are worse off being stochastically dominated by the aforementioned types (see panels 1 and 2 of Figures 1 and 2). Under $t = 0.5$ and $f = 1$, isolated effective altruists used to dominate deontological altruists and zero-intelligence agents; in polymorphic societies, they are dominated (compare panel 3 of Figures 1 and 2). Deontological altruists passing half of their endowment feature CDF similar to zero-intelligence altruists (see panel 3 of Figure 2). Finally, zero-intelligence agents stochastically dominate pathologically altruistic effective and deontological agents; however, deontological altruists perform better (see panel 4 of Figure 2).

The trade-off between allocative efficiency and equality becomes apparent when analysing the polymorphic group as a whole, rather than focusing on the individual types (Appendix B, Figure B2). The society of 'self-neutral' effective and deontological altruists (i.e. agents featuring $f = 1$ and $t = 0.5$) exhibits the smallest increase in wealth inequality. At the same time, pathological altruism (i.e. $f = 0.25$ and $t = 0.7$) implies a higher level of cumulative wealth.

In the next step, we assess the effect of deontological agents' giving heuristics while assuming that effective altruists are concerned with welfare maximisation solely (implying that $f = 1$). The results are displayed in Figure 3. Utterly selfish deontological altruists passing only 0.01 of their endowment perform better than welfare-maximising effective altruists. The result is not surprising because selfishness represents a 'predatory' strategy (giving almost nothing, yet receiving donations from others). Yet, even zero-intelligence agents perform better than effective altruists (see panel 1). The outcome remains similar after increasing the proportion of endowment donated by deontological altruists to $t = 0.2$ (panel 2). Deontological altruists donating half of their endowment perform similarly to zero-intelligence agents, but effective altruists' position remains inferior (panel 3). In the case of pathological deontological altruism ($t = 0.7$), approximately half of the effective altruists still perform worse than deontological altruists, although zero-intelligence agents dominate both types (panel 4).

For welfare-maximising effective altruists, the trade-off between efficiency and inequality in the whole society is not as sound as in the previously discussed simulation (see Figure B3 in Appendix B). Although the scenario of $t = 0.5$ and $f = 1$ still produces the most egalitarian outcome (i.e. the outcome associated with the slightest change in the Gini coefficient), the CDF is mainly located on the right-hand side of the CDFs corresponding to the remaining scenarios. Therefore, the society of 'self-neutral' deontological and effective altruists outperforms the societies consisting of 'self-neutral' (welfare-maximising) effective altruists and more selfish or more generous deontological altruists in terms of wealth and equality.

Finally, attempting to mirror the real-life example of deontological giving – the practice of tithing – we explore the societies where all the deontological altruists donate 0.1 of their endowment (see Figure 4). Relatively selfish deontological altruists outperform pathologically altruistic effective agents

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

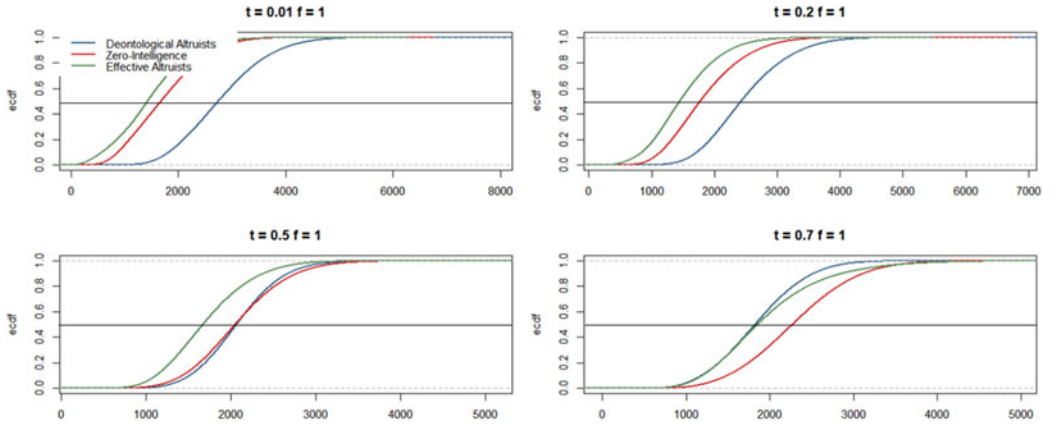

Figure 3. Wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming $f = 1$ for different values of t parameter.

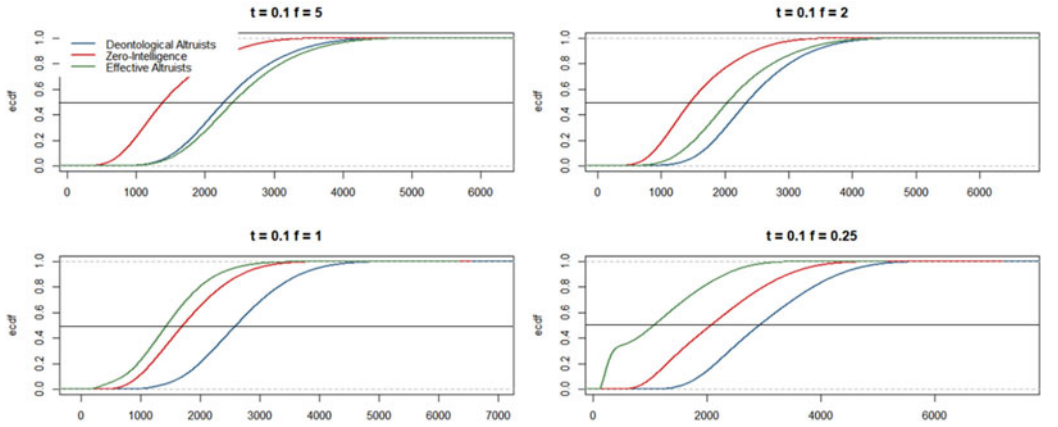

Figure 4. Wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming $t = 0.1$ for different values of f parameter.

(panel 4), which should not come as a surprise. However, in such a setting, a substantially higher degree of selfishness among effective altruists provides a minor advantage: although selfish effective altruists dominate deontological altruists, the distance between the CDFs is barely visible (panel 1).

Society as a whole (see Figure B4 in Appendix B) experiences the smallest increase in the Gini coefficient when both effective and deontological altruists are relatively selfish – i.e. when $t = 0.1$ and $f = 2$. Interestingly, a relatively low level of wealth inequality is achieved without considerable marginal wealth redistribution (which would require more generous deontological and effectively altruistic agents). Moreover, such a society stochastically dominates the population with welfare-maximising effective altruists (i.e. featuring $f = 1$).

Conclusions

Effective altruism – a paradigm that proclaims a pragmatic attitude to charitable giving – is becoming increasingly popular. Although it would be a mistake to equalise Benthamite utilitarianism and effective altruism, the prominent features of the latter – teleological orientation and emphasis on the 'greatest good' (or allocative efficiency) – have been inherited from the former. A society consisting of individuals

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

concerned with aggregate welfare maximisation should accumulate the greatest wealth. The results of our simulations based on the dictator game and incorporating effectively altruistic, deontologically altruistic, and zero-intelligence agents in isolation confirm this intuition. Although welfare-maximising (or 'self-neutral') agents perform the best, more (or less) selfish effective altruists still dominate the remaining groups. At the same time, the isolated populations of effective altruists experience the greatest increase in wealth inequality, whereas the deontological pattern of giving partially neutralises the initial disproportionate endowment distribution. At the same time, assuming the uniform initial endowment distribution, isolated welfare-maximising effective altruists outperform regarding the total wealth aggregated and wealth equality. Our results provide some additional evidence for the well-known equality-efficiency trade-off. The utilitarian pattern of giving might lead to the utopian outcome – a society featuring equality and allocative efficiency. Yet, this might happen solely on rare occasions, such as under the perfect equality of opportunities. In a more realistic setting, allocating resources following the maximum marginal benefit principle leads to a growing wealth disparity.

However, the objectives of the present paper go beyond demonstrating the anti-egalitarian implications of utilitarianism. Intuitively, social arrangements developed in evolutionary selection should deal with uncertainty and ensure resilience and sustainability better than norms being the product of deliberate design, such as effective altruism. This conjunction is tested by analysing the interaction between effective altruists, deontological altruists, and zero-intelligence agents in polymorphic societies under various settings. Even the most selfish effective altruists (i.e. the agents following the 'predatory' strategy) have only a minor advantage over deontological altruists. In contrast, for agents following the unconditional giving strategy, it is much easier to achieve stochastic domination over other types regarding wealth distribution. The deontologically altruistic strategy might underperform in terms of efficiency; yet it facilitates resilience in the polymorphic societies.

It should also be noted that in the real-world setting, deontological altruism heuristics constantly evolve on the macro- and the micro-levels. Individuals change interaction strategies depending on the signals from the external environment, and they can even develop complex behavioural patterns spontaneously (i.e. not requiring any intervention from an end-directed social planner). The simple simulation setting we utilised cannot reflect the whole complexity of deontological altruism. Therefore, if deontological altruists with no ability to learn perform better than effective altruists in most instances, then communities of actual human beings – featuring a proper sense of sociability, ability to learn, and come up with new social exchange rules – should be even more robust and resilient. We are convinced that the 'epistemological issues' of effective altruism are far less important than its fundamental flaw: in a world of radical uncertainty, social exchange patterns subject to evolutionary selection are far more sustainable and resilient than any attempt at meticulous 'social planning'.

Overall, effective altruism is an ideology worth popularising in the ideal world with perfect equality of opportunities, uniform attitudes towards giving, and no radical uncertainty. Yet, we live and interact in a much more complex setting. The paradox of effective altruism sounds as follows: although it strives to achieve the greatest good for humanity in the long-run perspective, the society of effective altruists appears to be the most anti-egalitarian and the least robust.

References

Alesina A. and Giuliano P. (2015). Culture and institutions. *Journal of Economic Literature* **53**(4), 898–944. https://doi.org/10.1257/jel.53.4.898

Andreoni J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal* **100**(401), 464–477.

Andreoni J. and Miller J. (2002). Giving according to GARP. *Econometrica* **70**(2), 737–753.

Aristotle. (2014). Cambridge texts in the history of philosophy. In Crisp R. (ed.), *Aristotle: Nicomachean Ethics*, 2nd edn. Cambridge: Cambridge University Press.

Arrow K.J. (1951). *Social Choice and Individual Values*. New York: Wiley.

Arrow K.J. (1997). The functions of social choice theory. In Arrow K.J., Sen A. and Suzumura K. (eds), *Social Choice Re-Examined*. London: Palgrave Macmillan, pp. 3–9.

Axelrod R. (1987). The evolution of strategies in the iterated prisoner's dilemma. In Davis L. (ed.), *Genetic Algorithms and Simulated Annealing*. London: Pitman, pp. 32–41.
Bentham J. (1823). *An Introduction to the Principles of Morals and Legislation*. London: Printed for W. Pickering. Jonathan Bennett. Available at https://www.earlymoderntexts.com/assets/pdfs/bentham1780.pdf
Bentham J. (1834a). Deontology. In Rosen F. (ed.), *The Collected Works of Jeremy Bentham*. Oxford: Oxford University Press, pp. 117–283.
Bentham J. (1834b). Article on utilitarianism. In Rosen F. (ed.), *The Collected Works of Jeremy Bentham*. Oxford: Oxford University Press, pp. 283–318.
Brandt R.B. (1979). *A Theory of the Good and the Right*. Oxford: Oxford University Press.
Colander D. (2000). The death of neoclassical economics. *Journal of the History of Economic Thought* **22**(02), 127–143. https://doi.org/10.1080/10427710050025330
Côté N. and Steuwer B. (2023). Better vaguely right than precisely wrong in effective altruism: the problem of marginalism. *Economics & Philosophy* **39**(1), 152–169. https://doi.org/10.1017/S0266267122000062
Elder M. and Fischer B. (2017). Focus on fish: a call to effective altruists. *Essays in Philosophy* **18**(1), 107–129. https://doi.org/10.7710/1526-0569.1567
Gabriel I. (2016). Effective altruism and its critics. *Journal of Applied Philosophy* **34**(4), 457–473. https://doi.org/10.1111/japp.12176
Galanis G. and Veneziani R. (2022). Behavioural utilitarianism and distributive justice. *Economics Letters* **215**, 110488. https://doi.org/10.1016/j.econlet.2022.110488.
Genç M., Knowles S. and Sullivan T. (2020). In search of effective altruists. *Applied Economics* **53**(7), 805–819. https://doi.org/10.1080/00036846.2020.1814947
Gill A. (2020). The comparative endurance and efficiency of religion: a public choice perspective. *Public Choice* 313–334. https://doi.org/10.1007/s11127-020-00842-1.
Gill A. and Thomas M.D. (2023). The dynamic efficiency of gifting. *Journal of Institutional Economics* **19**, 70–85.
Goodin R.E. (1995). *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.
Goodin R.E. (1998). Public service utilitarianism as a role responsibility. *Utilitas* **10**(3), 320–336. https://doi.org/10.1017/S0953820800006245
Haidt J. (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York: Pantheon Books.
Hammond P.J. (1991). Interpersonal comparisons of utility: why and how they are and should be made. In Elster J. and Romer J.E. (eds), *Interpersonal Comparisons of Well-Being*. New York: Cambridge University Press, pp. 200–254. https://doi.org/10.1017/cbo9781139172387.008
Hardt Ł. (2020). Utylitaryzm, deontologia i etyka cnót: zbieżne czy przeciwstawne fundamenty etyczne ekonomii? *Ekonomista* **2020**, 249–265.
Harsanyi J.C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* **63**(4), 309–321. https://doi.org/10.1086/257678
Harsanyi J.C. (1978). Decision theory and utilitarian ethics. *The American Economic Review* **68**(2), 223–228.
Hayek F.A. (1944). *The Road to Serfdom*. Chicago: University of Chicago Press.
Hayek F.A. (1945). Individualism: true and false. In Caldwell B. (ed.), *Studies on the Abuse and Decline of Reason*. Carmel: Liberty Fund, Inc., pp. 46–74.
Hillinger C. (2004). Utilitarian collective choice and voting. *SSRN Electronic Journal*, http://dx.doi.org/10.2139/ssrn.637521
Imhof L.A., Fudenberg D. and Nowak M.A. (2007). Tit-for-tat or win-stay, lose-shift? *Journal of Theoretical Biology* **247**(3), 574–580.
Johansson-Stenman O. (1998). On the problematic link between fundamental ethics and economic policy recommendations. *Journal of Economic Methodology* **5**(2), 263–297. https://doi.org/10.1080/13501789800000016
Jones B.L. and Unsworth R.K.F. (2020). The perverse fisheries consequences of mosquito net malaria prophylaxis in East Africa. *Ambio* **49**, 1257–1267. https://10.1007/s13280-019-01280-0.
Kant I. (1785). Fundamental principles of the metaphysics of morals. Retrieved from the Project Gutenberg, 2018. Available at https://www.gutenberg.org/files/5682/5682-h/5682-h.htm
Keynes J.M. (1921). *Treatise on Probability*. London: Macmillan.
Kopczewski T. and Okhrimenko I. (2019). Can *Homo economicus* be an altruist? A classroom experimental method. *International Review of Economics Education* **32**, 100167. https://doi.org/10.1016/j.iree.2019.100167.
Kropotkin P. (1904). *Mutual Aid*. London: Penguin Classics.
List C. (2003). Are interpersonal comparisons of utility indeterminate? *Erkenntnis* **58**, 229–260.
Little M.D. (2002). *Ethics, Economics, and Politics: Principles of Public Policy*. Oxford: Oxford University Press.
Lomborg B. (1996). Nucleus and shield: the evolution of social structure in the iterated prisoner's dilemma. *American Sociological Review* **61**(2), 278–307. https://doi.org/10.2307/2096335
MacAskill W. (2018). Understanding effective altruism and its challenges. In Boonin D. (ed.), *The Palgrave Handbook of Philosophy and Public Policy*. Cham: Palgrave Macmillan, pp. 441–453. https://doi.org/10.1007/978-3-319-93907-0_34.
MacAskill W. (2022). *What We Owe the Future*. New York: Basic Books.
McNaughton D. and Rawling P. (2007). Deontology. In Copp D. (ed.), *The Oxford Handbook of Ethical Theory, Oxford Handbooks*, Online Edn. Oxford: Oxford Academic, pp. 424–458. https://doi.org/10.1093/oxfordhb/9780195325911.003.0016

Mill J.S. (1863). *Utilitarianism*. London: Parker, son, and Bourn. Batoche Books Kitchener.
Myerson R.B. (1983). Bayesian equilibrium and incentive compatibility: an introduction. In Hurwicz L., Schmeidler D. and Sonnenschein H. (eds), *Discussion Papers 548*. Northwestern University, Center for Mathematical Studies in Economics and Management Science.
Nowak M. and Sigmund K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58. https://doi.org/10.1038/364056a0
Oakley B., Knafo A., Madhavan G. and Wilson D.S. (2011). *Pathological Altruism*. Oxford: Oxford University Press.
Rawls J. (1971). *A Theory of Justice*, Revised Edn. Cambridge: Harvard University Press.
Riley J. (2008). Utilitarianism and economic theory. In Durlauf S.N. and Blume L.E. (eds), *The New Palgrave Dictionary of Economics*, 2nd edn. London: Palgrave Macmillan, pp. 6923–6933. https://doi.org/10.1057/978-1-349-95121-5_2052-1.
Robbins L. (1938). Interpersonal comparisons of utility: a comment. *Economic Journal* **48**, 635–641. https://doi.org/10.2307/2225051
Russell, B. (1946). *A History of Western Philosophy*. Berkeley: Counterpoint, 1984.
Scanlon T.M. (1975). *Preference and urgency*. *Journal of Philosophy* **72**, 655–669. https://doi.org/10.2307/2024630
Sen A. (1972). Utilitarianism and inequality. Economic and Political Weekly. Available at https://ia903202.us.archive.org/6/items/UtalitarianismAndInequality-AmartyaSen/UtilitarianismAndInequality.pdf
Singer P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs* **1**(3), 229–243.
Singer P. (2009). *The Life You can Save*. New York: Random House.
Stark O., Kobus M. and Jakubek M. (2012). A concern about low relative income, and the alignment of utilitarianism with egalitarianism. *Economics Letters* **114**(3), 235–238. https://doi.org/10.1016/j.econlet.2011.10.024
Sugden R. (2018). *The Community of Advantage*. Oxford: Oxford University Press.
Sunder S. (1993). Allocative efficiency of markets with zero-intelligence traders: market as a partial substitute for individual rationality. *Journal of Political Economy* **101**(1), 119–137.
Sunstein C.R. (2014). Is deontology a heuristic? On psychology, neuroscience, ethics, and law. *Iyyun: The Jerusalem Philosophical Quarterly* **63**, 81–103. http://www.jstor.org/stable/23685973
van Staveren I. (2001). *The Values of Economics: An Aristotelian Perspective*. London: Routledge.
Wahl L.M. and Nowak M.A. (1999). The continuous prisoner's dilemma: II. Linear reactive strategies with noise. *Journal of Theoretical Biology* **200**(3), 323–338. https://doi.org/10.1006/jtbi.1999.0997
White M.D. (2009). In defense of deontology and Kant: a reply to van Staveren. *Review of Political Economy* **21**(2), 299–307. https://doi.org/10.1080/09538250902834103

Appendix A: *Giving according to GARP* experimental setting and prior results

Giving according to GARP experimental framework (Andreoni and Miller, 2002) is a modified version of the well-known dictator game. In the original version, the game consists of several rounds in which agents decide what proportion of their endowment to donate under different trade-off values (which are interpreted as 'prices' of altruism). The results are then used to analyse whether altruistic preferences are consistent with the GARP or rational in the neoclassical constructivist sense. Simulation in Kopczewski and Okhrimenko (2019) mirrors the experimental setting, although the research objectives are radically different: the experimental design is treated as a tool to incorporate the distinctions between consequentialist and deontological ethics into consumer theory.

The game consists of eight rounds (see Table A1). In each round, players receive some tokens, which serve as the initial endowment. Participants should decide how many tokens to hold (π_s) or pass to other players (π_o) depending on the hold and pass value. For instance, in the first round, each participant receives 40 tokens, the hold value is 3, and the pass value is 1. If a participant decides to keep 20 tokens and pass 20 tokens, the participant receives $20 \times 3 = 60$ tokens, and the other randomly selected player receives $20 \times 1 = 20$ tokens. All the choices must satisfy the budget constraint; with modified values of the endowment and the hold/pass values, there are eight budget constraints in total (see Figure A1).

In the game setting described above, 'payment to self' and 'payment to others' serve as broad economic goods, denoting selfish self-optimisation and altruism. Altruism serves as an economic good because it is associated with costs and positive marginal utility (see Andreoni, 1990). Using GARP, Andreoni and Miller (2002) verify whether altruistic preferences constitute a well-behaved utility function; moreover, the authors explore the most common forms of GARP-consistent altruistic preferences. In Kopczewski and Okhrimenko (2019), the *Giving according to GARP* (Andreoni and Miller, 2002) framework is used to discuss two ethical imperatives: teleological and deontological altruism. Teleological (or consequentialist) altruists seek to maximise social welfare. The greatest happiness principle governs effective altruists: if (and only if) the pass value is greater than the hold value, effective altruists should pass their endowments. Hold and pass values (see Table A1) can be interpreted as the marginal utility of players and their opponents associated with a token. The altruistic preferences of effective altruists are tested for rationality or consistency with the weak axiom of revealed preferences (WARP) and the strong axiom of revealed preferences (SARP). Figure A2 visualises effective altruists' choices between paying themselves ('Self' on the graph) and paying others under various budget constraints (see Table A1 and Figure A1). The graph on the left shows no violation of WARP (indicating that altruistic preferences are complete); moreover, as SARP is not violated (see graph on the right), effective altruists' preferences are also transitive.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

Table A1. Allocation choices

Round	Token endowment	Hold value	Pass value	p_s	p_o	Budget constraints
1	40	3	1	0.33	1	$0.33\pi_s + 1\pi_o = 40$
2	40	1	3	1	0.33	$1\pi_s + 0.33\pi_o = 40$
3	60	2	1	0.5	1	$0.5\pi_s + 1\pi_o = 60$
4	60	1	2	1	0.5	$1\pi_s + 0.5\pi_o = 60$
5	75	2	1	0.5	1	$0.5\pi_s + 1\pi_o = 75$
6	75	1	2	1	0.5	$1\pi_s + 0.5\pi_o = 75$
7	60	1	1	1	1	$1\pi_s + 1\pi_o = 60$
8	100	1	1	1	1	$1\pi_s + 1\pi_o = 100$

Source: Andreoni and Miller (2002: 740–741).

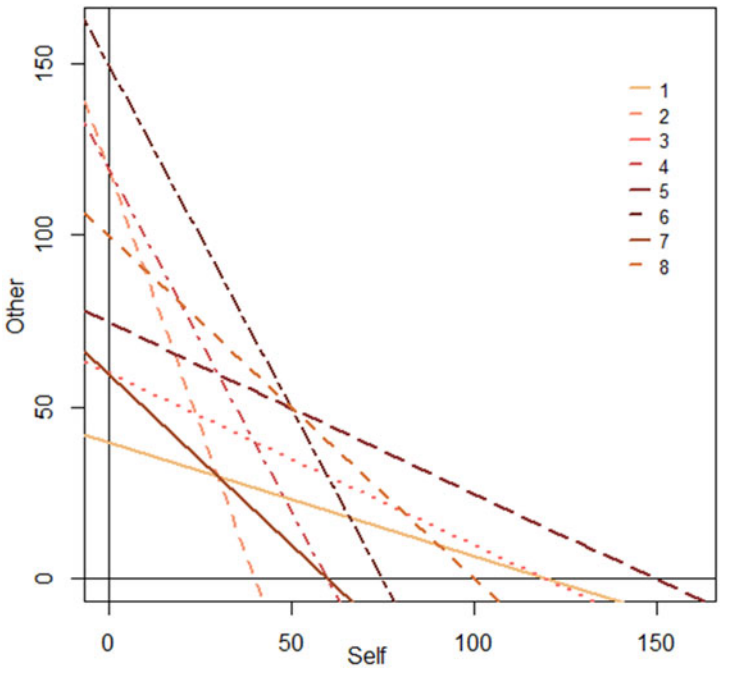

Figure A1. Budget constraints.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

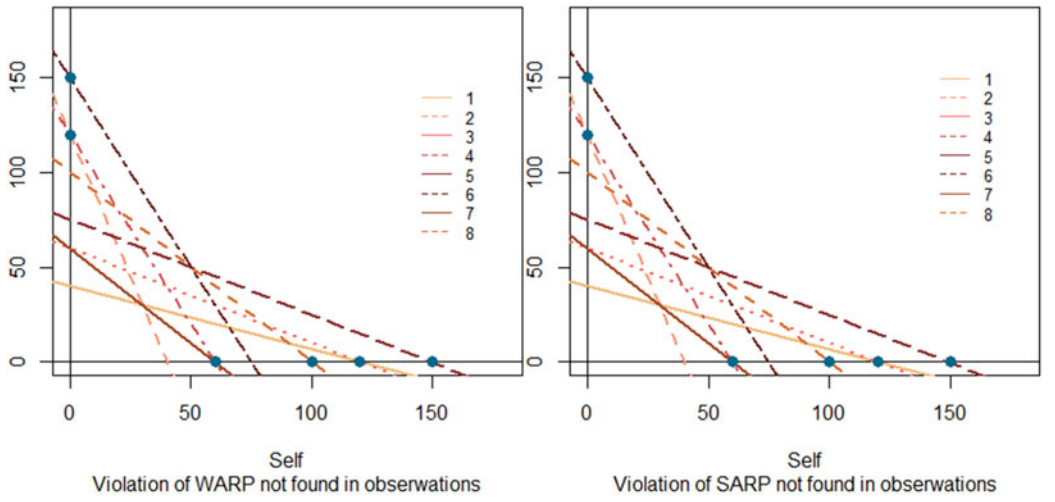

Figure A2. Testing effective altruism for consistency with WARP and SARP.
Source: Kopczewski and Okhrimenko (2019: 9).

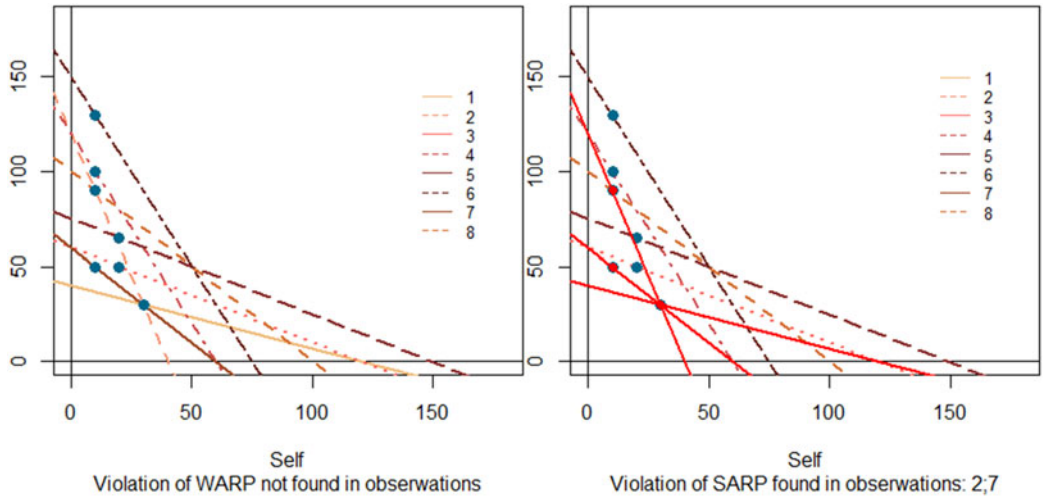

Figure A3. Testing deontological altruism for consistency with WARP and SARP.
Source: Kopczewski and Okhrimenko (2019: 10).

Unlike effective altruists, deontological altruists are guided by heuristics. Let us assume that deontological altruists pass the same proportion of tokens each round, regardless of the ratio of the pass and hold values (in line with the social heuristic logic described in the paper). In rounds 2, 4, and 6, society is better off because the pass value exceeds the hold value. Although deontological altruism would lead to incremental improvement, it will not lead to perfect allocative efficiency – this can only occur if all tokens are passed. In rounds 7 and 8, social welfare is not affected. Finally, in rounds 1, 3, and 5, deontological altruism would distort social utility: passing tokens when the hold value is greater than the pass value implies a waste of resources. Furthermore, suppose we stick to the substantive notion of rationality (i.e. consistency with optimisation principles). In this case, deontological altruism is irrational because it violates SARP (see Figure A3, which shows that the deontological pattern of giving violates SARP).

In summary, from a constructivist perspective, giving according to effective altruism is rational and leads to aggregate utility maximisation, in contrast to heuristic giving. Therefore, from the perspective of mainstream economics, effective altruism should be regarded as the most desirable pattern of social exchange.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

Appendix B: Selected figures

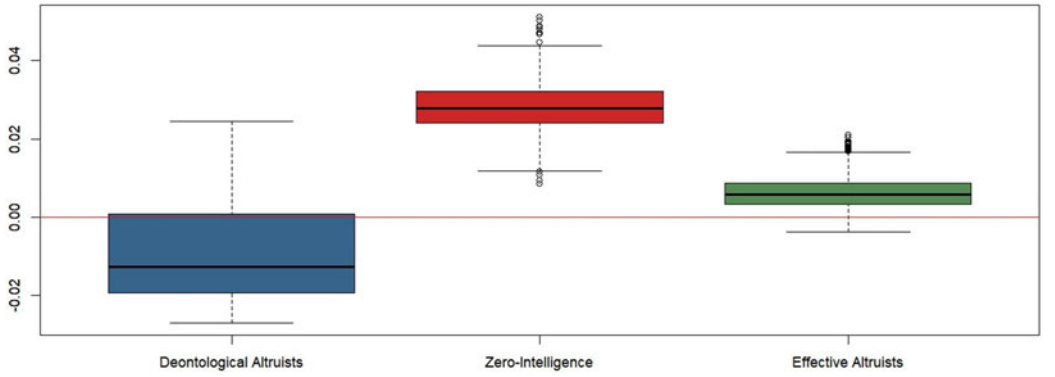

Figure B1. Boxplot of changes in Gini coefficients after each iteration in monomorphic societies of deontological altruists, effective altruists, and zero-intelligence altruists.

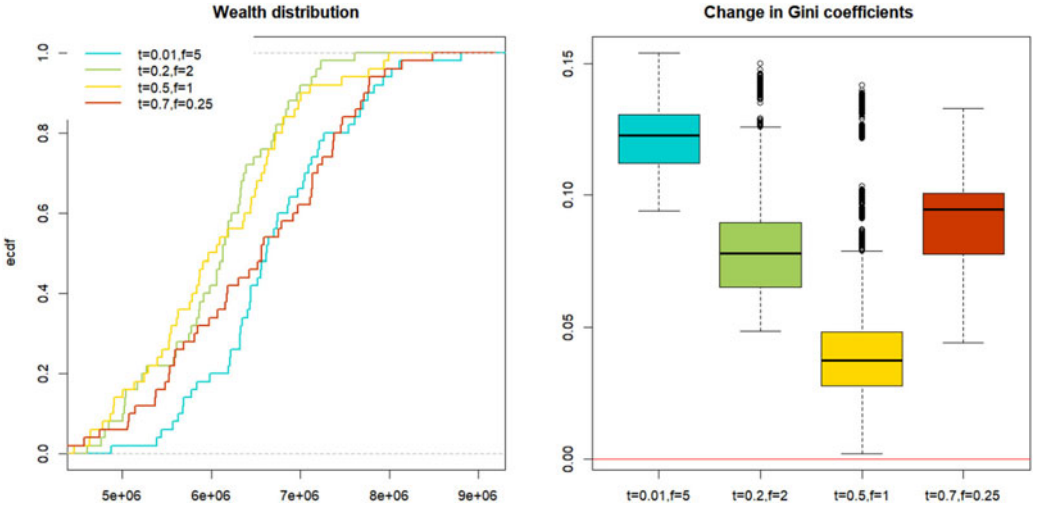

Figure B2. Aggregate wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming different values of *t* and *f* parameters.

https://doi.org/10.1017/S1744137424000146 Published online by Cambridge University Press

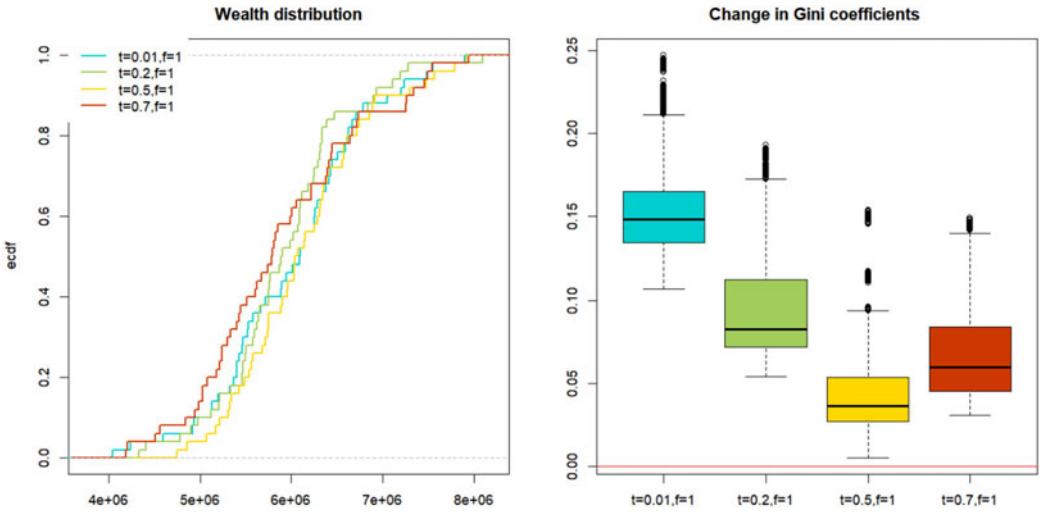

Figure B3. Aggregate wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming $f = 1$ for different values of t parameter.

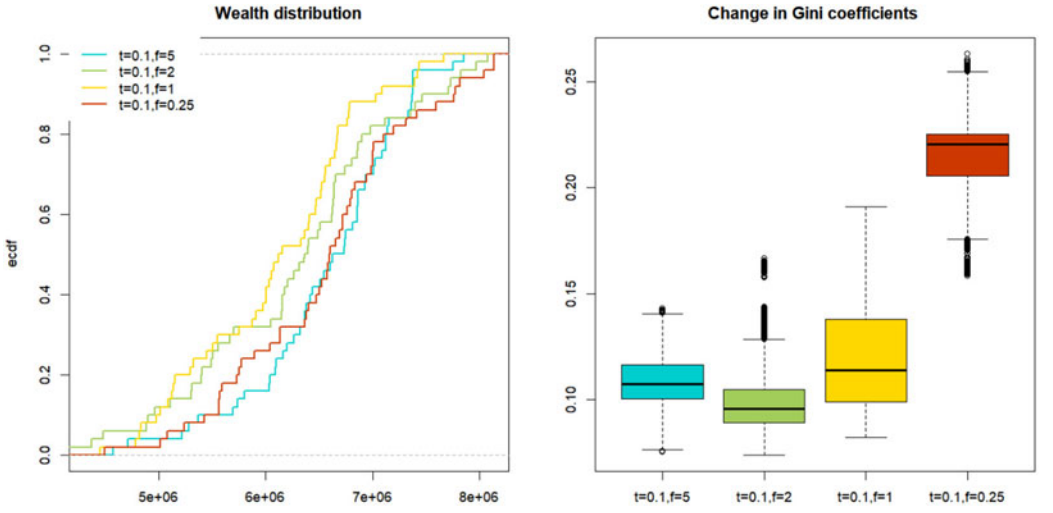

Figure B4. Aggregate wealth distributions in polymorphic societies combining deontological altruists, effective altruists, and zero-intelligence altruists in equal proportions assuming $t = 0.1$ for different values of f parameter.

Cite this article: Kopczewski T. and Okhrimenko I. (2024). The paradox of effective altruism. *Journal of Institutional Economics* **20**, e41, 1–18. https://doi.org/10.1017/S1744137424000146