Corrupting Effectiveness: Utilitarianism and Moral Impartiality toward Future Persons in Pragmatic Evaluation of Altruistic Interventions.

By Peter Scheyer[1]

Introduction: Effective Altruism and the Far Future Trillions Argument

In recent years billions of philanthropic dollars have been deployed through a movement and philosophy known as Effective Altruism, notably through the organizations Open Philanthropy, GiveWell, Good Ventures, and the over 3,200 persons taking the 'Giving What We Can' pledge to limit their personal income and donate the remainder to charity[2]. Effective Altruism, or EA, explicitly aims to 'use evidence and reasoning to determine the most effective ways to benefit others.'[3]

Within the EA community there are competing viewpoints on how to evaluate effectiveness, which forms of evidence and reasoning are best, and several leading philosophers with their own adherents and value systems. The discussions surrounding the employment of these viewpoints can grow heated, with one commenter complaining that 'Effective Altruism markets itself as being centered on addressing global poverty, when in fact it is centered on manipulating people into believing in the imaginary AI doomsday.'[4]

This complaint highlights a major schism within EA. On one side is the empirical evaluation of existing altruistic interventions according to their own internal justifications - the attempts to judge the most effective way to solve current issues like global poverty, using hard data and statistics. On the other side is the a logically based rationalist effort to determine undervalued new interventions, bring them appropriate funding, and potentially solve issues which are ignored by mainstream altruism.

From within the community, these two parts of the current EA community are often considered two sides of the same coin. Extrapolations of techniques used to evaluate current interventions often provide the justification for attempting new interventions, and can highlight areas worth a closer examination or the allocation of more funding.

Over time, a single paradigm has come to dominate many of these extrapolations, leading to the complaint above. Nicholas Beckstead, a program officer at Open Philanthropy, laid out the argument underpinning this paradigm in a 2013 doctoral thesis titled 'On the Overwhelming Importance of Shaping the Far Future.'[5] The argument in this thesis, which we will call the Far

---

[1] With assistance from The Foresight Institute, Haley Madel, Allison Duettmann, Benjamin Hoffman, Benjamin Scheyer, Clayton Faits, the Centre for Effective Altruism, and his wife Jennifer Scheyer, as well as various individuals who offered glad ears and advice.

[2] (Matthews)

[3] (Centre for Effective Altruism)

[4] (mystery-babylon)

[5] (Beckstead)

Future Trillions Argument (FFTA), is central to many of the most heated discussions in how to best employ the billions of dollars of the Effective Altruism.

After careful consideration it is our reasoned opinion that the Far Future Trillions Argument and its resulting recommendations, while internally coherent and defensible, are not within the remit of a movement characterized by the use of evidence in determining the effectiveness of altruistic interventions. This overall conclusion is based on seven separate arguments.

The first argument is based on empirical evaluation of foreign aid interventions, and concludes that characteristics of the FFTA are inimical to effective interventions. Our second and third arguments conclude that the FFTA lacks certain characteristics of pragmatic models, and requires empirically fallacious methodological assumptions to connect far future outcomes with present interventions. Fourth and fifth, we argue that the purely rationalist, philosophical nature of the FFTA divorces it from evidence in a way that inherently foils the determination of the effectiveness of interventions based upon it. Our sixth and seventh arguments take issue with the ongoing privileged employ of the version of utilitarianism and moral impartiality toward future persons used in the FFTA and its offshoot justifications, asserting that the inclusion of a full and wider scope of normative justifications is necessary for pragmatic modeling.

The Far Future Trillions Argument

Nicholas Beckstead's version of the Far Future Trillions Argument (FFTA)[6] is:

Humanity may survive for millions, billions, or trillions of years.

If humanity may survive may survive for millions, billions, or trillions of years, then the expected value of the future is astronomically great.

Some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways.

If the expected value of the future is astronomically great and some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways, then from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

Therefore, from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

To connect the FFTA to specific interventions three steps are necessary. To determine what is best (in expectation), we can first, evaluate potential outcomes using utilitarianism. Second, we can assign an approximate likelihood of our actions contributing to these outcomes. Third, we can value interventions using arguments based on resulting statements. For example, 'the fact that an existential risk of x% means that the future has x% less expected value than it would if there were no risk.'[7]

The FFTA is shortened and restated practically by the Centre for Effective Altruism[8] (CEA Handbook 2017) as: The long-term future has enormous potential: our descendants could live for billions or trillions of years, and have very high-quality lives; It seems likely there are things we can

---

[6] (Beckstead)

[7] (Beckstead), p.6

[8] (Centre for Effective Altruism)

do today that will affect the long-term future in nonnegligible ways; Possible ways of shaping the long-term future are currently highly neglected by individuals and society; Given points 1 to 3 above, actions aimed at shaping the long-term future seem to have extremely high expected value[9], higher than any actions aiming for more near-term benefits.

For our purposes, eight clarifications of the far-future trillions argument are useful.

1. The FFTA includes reasoning from long timeframes.
2. The FFTA includes reasoning from sets of possible specific futures.
3. The FFTA includes reasoning from the well-being of future persons.
4. The FFTA asserts we should attempt to improve, or 'aid,' future persons.
5. The FFTA explicitly asserts utilitarianism as a doctrine.
6. The FFTA is philosophically based on rationalism, 'a belief or theory that opinions and actions should be based on reason and knowledge rather than on religious belief or emotional response.'
7. The FFTA requires hypothetical estimations of approximate EV to make consequentialist arguments.
8. The FFTA requires the use of hypothetical probabilities and outcomes.

First Argument: Aid interventions incorporating the FFTA in their design are less likely than other interventions to be effective.

It is held within the pragmatic consequentialist framework that the justification for an intervention is immaterial to the outcome. However, a large body of work suggests that this is not the case, and that the arguments and spirit in which an intervention is undertaken can have subtle yet immense impacts on its effectiveness.[10] For our purposes, aspects of the FFTA place concerning handicaps, negatively influencing the metrics and scale of interventions, changing whether an intervention is aiming at a specific hypothetical future or an improved circumstance for its targets, and determining the frequency and relevance to planning of interactions with the aided.

This argument requires four premises. First, effective is defined as 'successful in producing a desired or intended result.' Second, aid interventions including long timeframes in their design are less likely to be successful in producing desired or intended results.[11] [12] Third, Aid interventions including possible specific futures in their design are less likely to be successful in producing desired

---

[9] Expected value (EV) can be defined as 'The expected value is the sum of the value of each potential outcome multiplied by the probability of that outcome occurring.'

[10] James C. Scott's *Seeing Like a State: How Certain Schemes to Improve The Human Condition Have Failed,* is an important overview work of this topic, while the examples used in the rest of this section are less philosophical and more empirical. For similar perspectives specifically applied to foreign aid interventions Easterly provides valuable context. (Scott), (Easterly, The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good)

[11] 'With smaller interventions, more rigorous evaluation is available to address the counterfactual question.' (Easterly, The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good), p. 53

[12] The use of randomized timeframes in evaluating the effectiveness of aid has been studied carefully and demonstrates that any lengthy, specific timeframe is less likely to be effective than randomization in timeframe selection. (Popper)(Duflo, Glennerster and Kremer)

or intended results.[13] [14] [15] Fourth, aid interventions which do not include provisions for feedback from the aided are less likely to be successful in producing desired or intended results.[16] [17]

As noted in clarifications 1 and 2, The FFTA requires reasoning from long timeframes and reasoning from sets of possible specific futures. Through its assumptions of what constitutes the well-being of future persons in clarification 3, the FFTA does not include provisions for feedback from the aided. Inclusion of an argument in a design includes the reasoning of the argument in the design.

In conclusion, aid interventions including the FFTA in their design are less likely to be successful in producing desired or intended results than some which do not incorporate the FFTA in their design.

Second Argument: Models including the FFTA without discounting are not justified as pragmatic in considering the relative impact of comparable altruistic efforts on the future.

Discounting refers to determining the present value of a payment or a stream of payments that is to be received in the future. Discounting methods are required to first, factor comparative opportunity costs into the model, and second, to consider compounding returns on alternatives which benefit from reinvestment.[18] Opportunity costs refer to the loss of potential gain from other alternatives when one alternative is chosen. Compounded returns arise when an investment results in an increase in resources, which can then be added to the original sum invested. When the new total is reinvested the returns are said to be compounding, using a definition of compound as a verb meaning to 'calculate (interest) on previously accumulated interest.'

Modeling without discounting is only justified as pragmatic if either the timeframe modeled does not include alternatives of comparable outcome, or alternatives of comparable outcome cannot be modeled.[19] Comparable alternatives to FFTA interventions exist in altruistic endeavors, and these comparable alternatives can be modeled. In conclusion, models including the FFTA without discounting are not justified as pragmatic.

---

[13] Karl Popper asserted that 'utopian social engineering' is less effective than piecemeal democratic reform in *The Poverty of Historicism,* p.61. (Popper)

[14] Abhijit Banerjee has compiled a lengthy list of possible interventions which have been verified as cost-effective uses of foreign aid. Linking such interventions together as needed is building an intervention out of proven blocks, as opposed to a utopian scheme. (Banerjee)

[15] Pitfalls of centralized planning in banking. (Whittle and Kuraishi)

[16] Easterly, *The White Man's Burden,* 'Westerners: don't do things to or for other people without giving them a way to let you know – and hold you accountable for – what you have actually done to or for them.' p. 381 'Discard your patronizing confidence that you know how to solve other people's problems better than they do.' p.368 (Easterly, The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good)

[17] *See* (De Renzio),

[18] *See* Kruschwitz, L., & Loffler, A. (2006). *Discounted Cash Flow: A Theory of the Valuation of Firms,* . West Sussex: John Wiley & Sons Ltd.

Third Argument: Justifications based on the FFTA are empirically fallacious, and as such are directly not justified in the philosophical system implied by colloquial use of the term 'effective.'

Argument 1 states 'the use of effective as an adjective only explicitly includes consequentialism and empiricism.' Deductions based on specific hypothetical probabilities, outside of understood contexts such as games or certain areas of physics, are empirically fallacious.[20] [21] [22] Empirically fallacious deductions are not justified within empiricist philosophical systems. The term 'effective' colloquially mandates an empiricist philosophical system.

Definitional clarification 7.1 states the FFTA requires hypothetical estimations of approximate EV, including the use of hypothetical probabilities and outcomes, to make consequentialist assertions. The FFTA is empirically fallacious when making consequentialist assertions.

In conclusion, justifications from consequentialist assertions of the FFTA are empirically fallacious, and as such are directly not justified in the philosophical system implied by colloquial use of the term 'effective.'

Fourth Argument: It is reasonable to expect a philosophy and social movement called Effective Altruism (EA) to only include philosophical positions justified by the colloquial use of the adjective 'effective.'

Our argument is based on five premises. First, effective is defined as 'successful in producing a desired or intended result.'[23] Second, altruism is defined as 'the belief in or practice of disinterested and selfless concern for the well-being of others.' Third, Effective Altruism (EA) self-defines as 'a philosophy and social movement that uses evidence and reasoning to determine the most effective ways to benefit others.'[24] Fourth, colloquial is defined as 'used in ordinary or familiar conversation.' And Fifth, usage of a word or phrase is 'colloquial' if it adheres to its ordinary and/or dictionary definition and usage.

'Effective altruism' could be strictly defined as an attempt to be successful in producing a desired or intended result, where that result is the belief in or practice of disinterested and selfless concern for the well-being of others. The stated definition of EA and the colloquial definition of effective altruism are close enough to warrant a literal interpretation of the constituent terms. It is reasonable to expect a philosophy and social movement to only include philosophical positions justified directly by a colloquial interpretation of its constituent terms.

It is *not* reasonable to expect a philosophy and social movement to *not* only include philosophical positions justified by a colloquial interpretation of its constituent terms, unless the additional inclusion is justified in one of three ways: directly, by additional constituent terms; indirectly, by philosophical positions included in a colloquial interpretation of the philosophy and social movement's constituent terms; contextually, by a requirement for the inclusion of additional philosophical positions to prevent the philosophy and social movement from being nonsensical.

---

[20] (Taleb, Fooled by Randomness; The Hidden Role of Chance in Life and in the Markets)

[21] (Taleb, The Black Swan: The Impact of the Highly Improbable)

[22] (Popper)

[23] All definitions used are from Google Search, Google. 11 Oct 2018. Web. 11 Oct 2018

[24] (Centre for Effective Altruism)

Altruism is restricted by the use of the adjective 'effective' in this context to the philosophical justifications implicit in the term 'effective.' The phrase 'effective altruism' is not nonsensical, and does not contextually require additional philosophical positions.

In conclusion, it is reasonable to expect a philosophy and social movement called Effective Altruism (EA) to only include philosophical positions justified by the colloquial use of the adjective 'effective.'

Fifth Argument: The use of the adjective effective only explicitly asserts a philosophical system requiring consequentialism and empiricism.

This argument uses six premises.

1. Effective is defined as 'successful in producing a desired or intended result.'
2. Altruism is defined as 'the belief in or practice of disinterested and selfless concern for the well-being of others.'
3. Consequentialism is defined as 'the doctrine that the morality of an action is to be judged solely by its consequences.'
4. Pragmatic Ethics assert that 'norms, principles, and moral criteria are likely to be improved as a result of inquiry.'
5. Empiricism is defined as 'the theory that all knowledge is derived from sense-experience.'
6. Result is synonymous with consequence.

In evaluating whether altruism is effective the success of an action desired or intended to be altruistic is judged by its desired or intended result. The conjunction of the terms effective and altruism asserts the success of altruistic action is judged by its desired or intended result. Effective altruism therefore seeks to make altruism more effective, placing a moral judgment on altruistic endeavors based on their success in producing their desired or intended result.

To the extent that it confers moral judgments solely based on effectiveness, use of effective as an adjective requires an adherence to consequentialism by definition. The past tense of effective's 'successful in producing' component implies that judgments of effectiveness are derived from data. Use of the adjective effective therefore implies adherence to empirical principles of knowledge derivation. The conjunction of empiricism and consequentialism defines the colloquial meaning of the term 'effective' as adhering to pragmatic ethics. Pragmatic considerations may justify alternative philosophies, if subjected to inquiry prior to inclusion.

The definition of effective consists of only the above parts. In this context, altruism is restricted to the philosophical justifications implicit in the term 'effective.' In conclusion, the use of effective as an adjective only explicitly asserts a philosophical system requiring consequentialism and empiricism.

Sixth Argument: Utilitarianism does not deserve automatic inclusion or a privileged place among pragmatic consequentialist justifications.

To deserve automatic inclusion among justifications a normative philosophy must lack relevant critiques. Utilitarianism has relevant critiques. Empiricist critiques of Utilitarianism suggest it is particularly poor for planning altruistic endeavors.[25] Deontologist critiques include the lack of

[25] (Popper)

utilitarian provision for human rights.[26] Virtue ethical critiques note that utility satisfaction prioritizes the apparent preferences of persons without seeking to improve said persons[27], and thus ends in a lowest-common-denominator version of humanity.[28]

There are even utilitarian critiques of the version of utilitarianism used in this discourse, noting that it tends to be a quantified ex post facto rationalization of personal preferences rather than a serious attempt to maximize utility.[29] Among normative philosophies utilitarianism is particularly prone to these issues due to its use of algorithms and quantified abstractions, while other normative philosophies insist that normative answers require philosophically engaging with hard moral questions.[30]

In order to deserve automatic inclusion among justifications a normative philosophy must lack equally compelling alternatives. Compelling alternatives to utilitarianism exist. Deontological justifications have data and good track records in aid.[31] Virtue Ethical justifications also exist.[32]

In conclusion, utilitarianism does not deserve automatic inclusion or a privileged place among altruistic justifications.

Seventh Argument: Moral Impartiality toward Future Persons does not deserve automatic inclusion or a privileged place among pragmatic consequentialist justifications.

To deserve automatic inclusion among justifications a normative philosophy must lack relevant critiques. Moral impartiality toward future persons has relevant critiques, two of which I will present here. First, a common-sense perspective suggests partiality toward oneself is reasonable, possibly even necessary, in sustained altruism[33]. Second, strict moral impartiality toward *present* persons and entities is an extremely demanding form of consequentialism[34], and the future variant glosses over these demands as 'outweighed by the mass of future persons' without due consideration.

To deserve automatic inclusion among justifications a normative philosophy must lack equally compelling alternatives, and compelling alternatives to moral impartiality toward future

---

[26] (Kant)
[27] (Kant), also (Sandel, Justice: What's the Right Thing to Do?)
[28] (Mill), (Kant), (Sandel, Justice: What's the Right Thing to Do?)
[29] (Sandel, Justice: What's the Right Thing to Do?)
[30] (Sandel, What Money Can't Buy: The Moral Limits of Markets)
[31] (Easterly, The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good)
[32] Aquinas wrote extensively on the topic of Charity in the development of virtue, see Sherwin, Michael S. *By knowledge & by love: charity and knowledge in the moral theology of St. Thomas Aquinas.* CUA Press, 2005.
[33] The existence of the concept of supererogation requires the concept of a baseline amount of duty required of moral beings. Efforts in other areas, i.e. personal sustenance, involves a degree of personal partiality if they are considered mandatory, yet most altruists agree a certain amount of responsibility must be taken for one's own well-being, and that concomitant efforts are excusable.
[34] *See* Pinker, Steven. *Enlightenment now: The case for reason, science, humanism, and progress.* Penguin Books, 2019.

persons exist. Partiality toward present persons and interventions is preferred by pragmatic empiricists[35].

In conclusion, Moral Impartiality does not deserve automatic inclusion or a privileged place among altruistic justifications.

1. Bibliography

Banerjee, Abhijit V. "Making Aid Work: How to fight global poverty—Effectively." *Boston Review* (2006).

Beckstead, Nick. "ON THE OVERWHELMING IMPORTANCE OF SHAPING THE FAR FUTURE." May 2013. *Rutgers Library.* 26 Sep 2018. <https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/>.

Bostrom, Nick. "Existential Risk Prevention as Global Priority." *Global Policy* 4.1 (2013): 15-31.

—. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Centre for Effective Altruism. "The Centre for Effective Altruism's Effective Altruism Handbook." 2 May 2018. *Centre for Effective Altruism.* 26 Sep 2018. <https://assets.ctfassets.net/ohf186sfn6di/glbXAUtnb2QagqY88qy4s/f8da9e4617efb89c0f79bf592b3f7ecd/Effective_Altruism_Handbook.pdf>.

De Renzio, P. "Accountability dilemmas in foreign aid." Working Paper. Overseas Development Institute., 2016.

Duflo, Esther, Rachel Glennerster and Michael Kremer. "Using Randomization in Development Economics Research: A Toolkit." *MIT Department of Economics Working Paper No. 06-36* (2006): 89.

Easterly, William. *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*. New York: Penguin Group (USA), 2006.

—. "Think Again: Debt Relief." *Foreign Policy* (2001): 20.

International Monetary Fund. "THE USES AND ABUSES OF SOVEREIGN CREDIT RATINGS." October 2012.

Kant, Morals Immanuel. *Foundations of the Metaphysics of Morals* Translated by Lewis White Beck. Seven Masterpieces of Philosophy. Routledge, 2016. 285-336.

Kruschwitz, L., & Loffler, A. (2006). *Discounted Cash Flow: A Theory of the Valuation of Firms,* . West Sussex: John Wiley & Sons Ltd.

Mill, John Stuart, and Mary Warnock. *Utilitarianism and On Liberty: Including 'Essay on Bentham' and Selections from the Writings of Jeremy Bentham and John Austin*. Malden: Blackwell, 2003.

Popper, Karl. *The Poverty of Historicism*. London and New York: Routledge, 1957.

Sandel, Michael J. *Justice: What's the Right Thing to Do?* New York: Farrar, Straus and Giroux, 2009.

—. *What Money Can't Buy: The Moral Limits of Markets*. First Paperback Edition, 2013. New York: Farrar, Strauss and Giroux, 2012.

Scott, James C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition have Failed*. Binghamton: Vail-Ballou Press, 1998.

Taleb, Nassim Nicholas. *Fooled by Randomness; The Hidden Role of Chance in Life and in the Markets*. New York: Random House, 2005.

—. *The Black Swan: The Impact of the Highly Improbable*. New York: Random House, 2007.

35

Whittle, Dennis and Mari Kuraishi. "Competing with Central Planning: Marketplaces for International Aid." *Reinventing Foreign Aid*. Ed. William Easterly. 1. Vol. 1. Cambridge: The MIT Press, 2008.