‘PHILOSOPHISCHE’ KRITIK AM EFFEKTIVEN ALTRUISMUS

Anton Leicht

SmP-Seminar ‘Effektiver Altruismus’

INTRO & INTERESSENSKONFLIKTE

- Speaker: Anton Leicht
- Promotionsstudent in Philosophie zu normativen Grundlagen von KI-Regulierung
- Nebenbei Policy-Arbeit zu KI in Berlin
- Keine aktuelle Affiliation mit EA-Organisationen
- Keine aktuelle Finanzierung durch EA-Fördermittel
  - Sommer 2022: Dreimonatiges Forschungsprojekt bei EA-naher Organisation
- Kein "institutional Effective Altruist", aber nicht ohne Vorprägung

STRUKTUR

VORWORT

ÖFFENTLICHE KRITIK

- Fokus dieses Moduls: Kritik an den *Ideen und Konzepten* von EA.
- Aber: Es gab in letzter Zeit viel Kritik an *spezifischen Personen und Organisationen* im Kontext von EA. Die sollte nicht unerwähnt bleiben.

VORWORT

ÖFFENTLICHE KRITIK II

- Gemeinsamer Trend: Zu schnell, zu viel, zu einfach gemacht
- Fehlende Distanz zwischen persönlichen und professionellen Kontexten: Verständlich für kleine Grassroots-Bewegung, inakzeptabel für große Institution
- Fehlende Governance-Normen in der Organisationsgestaltung: Üblich für kleine Non-Profits, inakzeptabel für 1000-Kopf-Organisationen
- Unkritischer Umgang mit Fördermitteln und Förderern: Notwendig bei chronisch unterfinanzierten Nischen-NGO, inakzeptabel für Milliarden-USD-Projekte


U.S.

Effective Altruism Promises to Do Good Better. These Women Say It Has a Toxic Culture Of Sexual Harassment and Abuse


The good delusion: has effective altruism broken bad?

A group of young idealists wanted to live the most ethical lives possible. Now some wonder whether the movement they joined has lost its moral compass

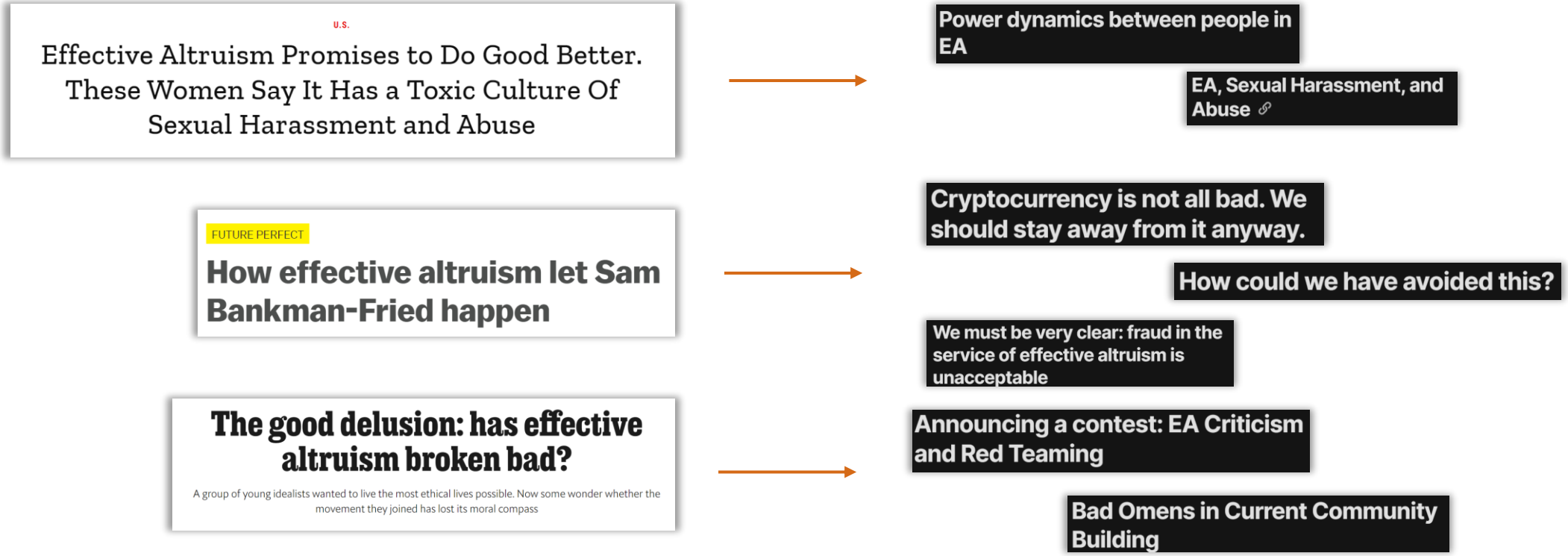

FUTURE PERFECT

How effective altruism let Sam Bankman-Fried happen

VORWORT

ÖFFENTLICHE KRITIK III

- Ein Lichtblick: Starkes Problembewusstsein und große Motivation zur Besserung
- Eine Einordnung:Viel Kritik ist konzentriert auf wenige Organisationen an der US-amerikanischen Westküste. Ein großer Teil der Organisationen ist nicht betroffen.



VORWORT

DREI SCHLÜSSELHYPOTHESEN

Man sollte im Leben möglichst effektiv Gutes tun

Effektiv Gutes zu tun, bedeutet, einer rigorosen Kosten-Nutzen-Abwägung zu folgen.

Die effektivsten Maßnahmen richten sich auf existenzielle Risiken.

Hypothese I

"Man sollte im Leben möglichst effektiv Gutes tun."

EFFEKTIVITÄT

KONSEQUENZIALISMUS

- Konsequenzialistische Ethik bewertet Handlungen nach ihren Konsequenzen.
- Effektivität ist vor Allem ein konsequentialistischer Begriff – Effektivität wird im Wert der Konsequenzen bemessen. Konsequenzen sind messbar, zählbar, 'stapelbar' – andere Güter vielleicht nicht.
- Aber Konsequenzialismus ist kontrovers. Viele Menschen sind überzeugt von…
- …Absoluten Pflichten und Geboten – z.B. einem religiösen Verhaltenskodex. Pflichterfüllung hat häufig keinen Effektivitätsmaßstab – entweder man entspricht den Pflichten oder eben nicht.
- …Politisch motivierten Prinzipien – z.B. "Earning to Give stützt eine ungerechte Gesellschaft". Wenn wir solchen Prinzipien folgen, ist individuelle Effektivität nicht der Maßstab.
- …Direkter gegenseitiger Verantwortung – z.B. einer prioritären Verpflichtung gegenüber der eigenen Familie.

EFFEKTIVITÄT

DEMANDINGNESS

Man kann immer *noch* mehr tun. Aber muss man?

- Möglichst effektiv Gutes zu tun, stellt sehr hohe Ansprüche und enthält keinen natürlichen 'Endpunkt' für Altruismus.
- Maximale Effektivität lässt keinen Raum für Supererogation, also den Gedanken: Wir sind moralisch verpflichtet, keinen Schaden anzurichten; Gutes zu tun, ist optional.
- Ein möglicher Ausweg: Einige Konsequenzialisten sind überzeugt von…
  - ➢ 'Kinship Consequentialism' – Wir sollten die eigene Familie, Gruppe, Nationalität, (Person) etc. priorisieren
  - ➢ 'Satisficing Consequentialism' – Wir müssen nicht maximal gute Konsequenzen bewirken, sondern nur hinreichend gute Konsequenzen.

Hypothese II

"Effektiv Gutes zu tun, bedeutet, einer rigorosen Kosten-Nutzen-Abwägung zu folgen."

KOSTEN-NUTZEN-RECHNUNGEN

RECHENSCHWIERIGKEITEN

Math is hard.

- Empirische & Moralische Unsicherheit / 'Cluelessness': Wir wissen sehr wenig über die Zukunft. Und wir wissen wenig darüber, was moralisch richtig ist. Diese Unsicherheit muss unsere Rechnung einbeziehen.
- Holistische Kosten-Nutzen-Rechnung, selbst zur groben Einordnung, ist ein umfangreiches und anspruchsvolles Unterfangen.
- Gleichzeitig: Viele bestehende Normen und Überzeugungen haben wichtige, schwierig nachvollziehbare Funktionen. Wenn wir selbst rechnen, lassen wir viel hart erarbeitetes gesellschaftliches Wissen zurück.
- Das heißt: Theoretisch kann man vielleicht die beste Karriere ausrechnen. Das heißt aber nicht, dass Kosten-Nutzen-Rechnung praktisch eine vollständige oder optimale Strategie ist.

KOSTEN-NUTZEN-RECHNUNGEN

FALLSTRICKE

Ein allmächtiges Wesen bietet euch ein Glücksspiel an. Mit einer Chance von 51% verdoppelt ihr den Wetteinsatz, mit einer Chance von 49% verliert ihr alles. Spielt ihr das Spiel? Wenn ja, wie oft?

- Der Erwartungswert des Spiels ist positiv. Naive Kosten-Nutzen-Rechnung rät, unendlich oft zu spielen – und wir verlieren alles.
- Sam Bankman-Fried hätte – und hat - das Spiel gespielt.
- Oberflächliche Implikation: Kosten-Nutzen-Rechnung gibt manchmal schlechten Rat.
- Aber: Wenn wir alle direkten und indirekten Effekte einpreisen, vielleicht auch nicht.
- Wichtigere Implikation: Durch die Auswahl von eingepreisten Effekten kann Kosten-Nutzen-Rechnung viele unterschiedliche Empfehlungen geben. Das birgt die Gefahr von Selbst- und Fremdbetrug.

Hypothese III

"Die effektivsten Massnahmen richten sich auf die Abwehr von existenziellen Risiken."

EXISTENZIELLE RISIKEN

WERT ZUKÜNFTIGEN LEBENS

Eine Frau mit Kinderwunsch geht zu ihrer Ärztin und erhält folgende Einschätzung: Würde sie heute schwanger werden, würde sie Zwillinge bekommen; aber die Schwangerschaft ginge mit Krankenhausaufenthalten und starken Schmerzen einher. Würde sie in sechs Monaten schwanger, verliefe die Schwangerschaft unproblematisch, aber statt Zwillingen bekäme sie nur ein Kind.

- Unsere Intuition sagt: Zukünftiges Leben auf Kosten von realem Wohlergehen zu schaffen, ist häufig nicht moralisch erforderlich.
- Das Argument für existenzielle Risiken sagt: Das Verringern von existenziellen Risiken ermöglicht das Leben vieler potenzieller zukünftiger Menschen.
- Aber: Das geschieht auf Kosten des Wohlergehens heutiger Menschen, denen wir stattdessen helfen könnten.
- Wie passt das zu unseren alltäglichen moralischen Intuitionen?

EXISTENZIELLE RISIKEN

DIE GRÖSSE DER ZUKUNFT

Das Argument für existenzielle Risiken enthält zwei wichtige Aussagen:

(1) Abwehr von heutigen existenziellen Risiken ermöglicht eine große und wertvolle Zukunft

(2) Technologischer Fortschritt bringt eine Reihe schwerwiegender existenzieller Risiken

- Aus (1) geht hervor: Der Erwartungswert der Abwehr existenzieller Risiken ist – selbst bei niedriger Eintrittswahrscheinlichkeit - groß, weil der Wert der Zukunft riesig ist.
- Aber (2) widerspricht: Durch viele existenzielle Risiken ist der Wert der Zukunft und damit der Erwartungswert der Abwehr einzelner Risiken gar nicht so hoch.
- Einfach gesagt: Wenn die Welt wahrscheinlich sowieso bald untergeht, ist es nicht unglaublich viel wert, sie heute zu retten.
- Restrisiken existenzieller Risiken reichen vielleicht nicht, um ihre Priorisierung zu motivieren.

TAKEAWAY?

EIN MODERATER ANSATZ

- ‘Robustness’: Viele klassische Karrieren und kurzfristigere Projekte sind in den meisten zukünftigen Welten gut. Indem man Robustness priorisiert, reduziert man die Chance, dass man am Ende nichts erreicht hat.
- Synergetische Interventionen: Es gibt viele Maßnahmen, die sowohl langfristige existenzielle Risiken minimieren als auch kurzfristigen Mehrwert in vielen Welten haben. Solche Karrieren enthalten sowohl den ambitionierten Gedanken des Effektiven Altruismus als auch viel der Sicherheit konventioneller Wege.
- Probleme treten vor Allem dann auf, wenn Abwägungen gegen allgemeine Normen oder gegen kurzfristige Interventionen getroffen werden. Solche Initiativen verdienen vorsichtige Betrachtung.

Q&A / DISKUSSION

AKADEMISCHE LITERATUR

Konsequenzialismus: Sinnott-Armstrong, W. (2023). Consequentialism. *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.)

Demandingness: Kagan, S. (1984). Does Consequentialism Demand Too Much? *Philosophy and Public Affairs 13*(3): 239-254.

Rechenschwierigkeiten: Lenman, J. (2000). Consequentialism and Cluelessness. *Philosophy and Public Affairs 29*(4): 342-370.

Fallstricke: Bernoulli, D. (1738). Exposition of a New Theory on the Measurement of Risk. *Econometrica 22*(1954): 23-36.

Wert zukünftigen Lebens: Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Größe der Zukunft: Thorstad, D. (2023). High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation. *Philosophy and Public Affairs 51*(4): 373-412.

SONSTIGE REFERENZEN

Demandingness: https://utilitarianism.net/objections-to-utilitarianism/demandingness/

Fallstricke: https://conversationswithtyler.com/episodes/sam-bankman-fried/

Existenzielle Risiken: https://ineffectivealtruismblog.com/