Description of primary metric [External]

The purpose of this document is to outline how we go about the assessment of our primary metric - expected career contribution based on Future Academy principles. This document is authored by Sebastian Schmidt (SS). Before the primary metric is described I want to make a couple of important points.

- This document attempts to assess to what extent we've served the world. This involves an assessment of the wonderful fellows who has participated in our programs along with our ability to enable them to make a difference in the world. This is **not** meant as an evaluation of peoples' worth nor as a definite score and we hope that no one feels offended by this.
- This is our current best *guess* for our default primary metric but impact evaluation **is extremely hard** and this isn't *the* model but rather *a* model. It's important that we don't Goodhart this metric.
- Please do not share this document with anyone without sufficiently high context and fidelity.

Here's what the document seeks to cover:

Introduction

We were inspired by Highly Engaged Effective Altruist[1] and 80,000hours' Impact-Adjusted Significant Plan Changes. We were dissatisfied with Highly Engaged Effective Altruist[2] as it doesn't reflect the heavy-tailed distribution of impact (it's 0 or 1) as well as overemphasizes the importance of engaging with the Effective Altruism community rather than doing good. We were somewhat satisfied with 80,000hours' impact-adjusted significant plan changes (IASPC). So, inspired by these metrics we created our own. It uses a mix of data in the form of surveys, observable outputs (e.g., the Impact Projects handed in by fellows, reported career strategies, and Forum posts), and personal interactions with fellows. We're stressing that evaluating the counterfactual impact is complex, and this metric will have important limitations too. It is, however, our best guess. **We call the metric counterfactual expected career contribution (CECC)**.

To give an understanding of the CECC metric, we'll give an example. Take an imaginary fellow, Alice. Before the intervention, based on our surveys and initial interactions, we expected that she may have an impactful career, but that she is unlikely to pursue a priority path based on IA principles. We rate her Expected Career Contribution (ECC) to be 2. After the program, based on surveys and interactions, we rate her as 10 (ECC) because we have seen that she's now applying for a full-time junior role in a priority path guided by impartial altruism. We also asked her (and ourselves) to what extent that change was due to IA and estimated that to be 10%. To get our final Counterfactual Expected Career Contribution (CECC) for Alice, we subtract her initial ECC score of 2 from her final score of 10 to get 8, then multiply that score by 0.1 to get the portion of the expected career contribution, which we believe we are responsible for. The final score is 0.8 CECC. As a formula: (10 (ECC after the program) - 2 (ECC before the program)) * 0.1 (our counterfactual influence) = 0.8 CECC. Below, you'll find details on how to evaluate ECC and our counterfactual influence.

We use the following ECC categories:

Category	Description

[1] A highly engaged effective altruist is a concept used by CEA. It includes the following components:

- Are using High-quality reasoning to determine which actions to take
- Are motivated by Impartially altruistic principles in deciding to take that action
- Stand a significant chance of becoming Leaders, thinkers, and philanthropists

[2] A highly engaged effective altruist is a concept used by CEA. It includes the following components:

- Are using High-quality reasoning to determine which actions to take
- Are motivated by Impartially altruistic principles in deciding to take that action
- Stand a significant chance of becoming Leaders, thinkers, and philanthropists

<0	They're on a career trajectory that we think is actively harmful. Clear red flags that could lead to bad outcomes for the world as well as the community. Something that feels off. E.g., optimizes too much for optics, very high ego, manipulation, excessive negligent or other dark-triade characteristics. Want to avoid Sam Bankmann-Fried situations and more extreme forms of malevolent actors.
0	Applied for Future Academy due to happenstance or seemed to not pursue something particularly promising with their career (based on our view of the world).
1	Sustained interest in similar programs. I.e., have a sense of them being on a trajectory to do something impactful but no concrete plans or contributions.
2	They're doing something that indicates they'll be able to do junior level contributions to priority path within months or a year but we're unsure about their interest in or ability to excel at a priority path (becoming a 100). This is similar to a "could go either way category".
10	They're contributing to a **junior** level impact option and/or have >10% of becoming a 100. This category requires a mix of *significant action* (great publications or junior level jobs) AND convincing case of their interest in and ability to excel at a priority path (career strategy and general impressiveness) based on IA principles.
20	Someone who **are doing a junior level** position based on IA principles and are on track to become 100.
100	They're contributing in a way that corresponds to a **moderately senior position** (10+ years of relevant work) in a promising path and/or have around 10% of becoming a 1000. Leader of a field that's valuable for the fabric of society even though not a priority path from us.
1000	They're among the very best in the world at what they do and pioneered important cause areas - in a priority career path.

Guide for evaluating the CECC

Below you'll find a simple description of how to do the ranking.

1. Get the data ready (surveys, etc.).
2. Prepare the spreadsheet.

3. At the beginning of the intervention, do a baseline assessment of the **expected career contribution based on IA principles (ECC)**
    a. Use the application form and do step 4.
    b. Assess "objective and observable outputs" and tentatively place them in power-law-like distributed categories (<0; 0; 1; 2; 10; 20; 100; 1000).
    c. Assess IA principles by loosely assessing the individual principles on a scale from 1-5 and then give a holistic assessment to adjust the category you placed them in.
4. At the end of the intervention, reassess the **expected career contribution based on IA principles (ECC).**
5. Assess our **counterfactual influence** on the fellow based on our assessment and self-assessed estimate in the survey.
6. Subtract the baseline ECC from the end ECC and multiply by our counter-factual influence to get **Counter-factual expected career contribution based on Impact Academy principles (CECC)** (our final impact eval score).

Expected career contribution based on Impact Academy principles

These are the proposed categories of impact measurements. They combine observable things such as the output of impact projects or specific job positions and our sense of whether they're guided by Impact Academy principles and no red flags.

Objective and observable outputs (rubric 1)

Here we're looking for promising outputs such as

- Blog posts on the EA forum or related forums (we trust that we'd hear/know about forum post outputs).
- Thesis on relevant topic (could be via the impact project)
- Relevant educational choice (LinkedIn and the survey).
- Their impact projects (final evaluations).
- Jobs and internships (assessed via survey, conversations, and LinkedIn).
- Donations (note, this is somewhat less important)
- Their stated career plans and cause prioritization (mix between the surveys, the impact project, and via interactions with them).

The purpose of this assessment is to explicate our reasoning and make sure that we check the relevant outputs in our assessments. Based on this, we already have a good idea of the final category we're likely to place them in (the expected career contribution based on FA principles).

Impact Academy principles (rubric 2)

Here we're using a holistic assessment of the following principles.

- Truth-seeking & high-quality reasoning (1-5)
- Morality (1-5)
- Development and well-being (1-5)
- Ability to excel (1-5)
- Red flags (0-3)

For an operationalization of these principles, please refer to this spreadsheet.

The purpose of this assessment is to explicitly think through some of the aspects of people we think are the most important as they'll guide peoples' trajectories and thus expected career contributions. This rubric is used to adjust the estimate above upwards or downwards.

These principles are assessed via interactions with them, things we've heard about them (e.g., via mentors), intuition, and the quality of the things above objective and observable things.

Expected career contribution based on IA principles (rubric 3)

Based on the objective and observable outputs and Impact Academy principles, we use a holistic assessment to place fellows into categories. This table gives an overview of the different categories of individuals we are using to give an ECC score. This score represents our best bet of the expected value of their contribution and doesn't include our counter-factual influence. For example, a person can start out as a 10 and still be a 10 after having participated in the program. Note, this is assessed at the beginning and end of the program.

Category	Description	Examples
<0	They're on a career trajectory that we think is actively harmful. Clear red flags that could lead to bad outcomes for the world as well as the community. Something that feels off. E.g., optimizes too much for optics, very high ego, manipulation, excessive negligent or other dark-triade characteristics. Want to avoid Sam Bankmann-Fried situations and more extreme forms of malevolent actors.	Redacted

0	Applied for Future Academy due to happenstance or seemed to not pursue something particularly promising with their career (based on our view of the world).	
1	Sustained interest in similar programs. I.e., have a sense of them being on a trajectory to do something impactful but no concrete plans or contributions.	
2	They're doing something that indicates they'll be able to do junior level contributions to priority path within months or a year but we're unsure about their interest in or ability to excel at a priority path (becoming a 100). This is similar to a "could go either way category".	
10	They're contributing to a **junior** level impact option and/or have >10% of becoming a 100. This category requires a mix of *significant action* (great publications or junior level jobs) AND convincing case of their interest in and ability to excel at a priority path (career strategy and general impressiveness) based on IA principles.	
20	Someone who **are doing a junior level** position based on IA principles and are on track to become 100.	
100	They're contributing in a way that corresponds to a **moderately senior position** (10+ years of relevant work) in a promising path and/or have	

	around 10% of becoming a 1000. Leader of a field that's valuable for the fabric of society even though not a priority path from us.	
1000	They're among the very best in the world at what they do and pioneered important cause areas - in a priority career path.	

Counter-factual expected career contribution based on Impact Academy principles (CECC)

Once we have a beginning and end ECC, we adjust for our counter-factual influence. This is because the changes that happened during our program are unlikely to be entirely due to us (e.g., at least one fellow did an EA-related residency which they'd likely have done regardless). Therefore we subtract the ECC estimate by our estimate of where they were when they started the program (the baseline) and multiplying that change by the counterfactual influence with these values

- 0: Not at all due to IA (<2%)
- 0.1: Somewhat due to IA (~10%)
- 0.5: Substantially due to IA (~50%)
- 1: Completely due to IA (~100%)

Example: (10 (ECC) - 2 (baseline))*0.5 (our influence)=4 CECC.

Note, we expect most impact can only be observed after a year or two and that this is just very hard and it's important to remain humble. 80,000 Hours has found "that there are long delays between when people engage with our programmes and when they make and get tracked as plan changes". The median delay (as of 2019) is two years, and only one has ever been tracked within one year." (resource).