RAISE status report
March 2018

Summary

RAISE was established in July 2017 as an initiative to improve the pipeline for AI safety research. Our mission is to unlock the capacity of aspiring AI safety researchers by creating an online course with high-quality content that's clear, distilled, and open to everyone. Milestones so far include building a study group and creating a first lecture video with assignments. Current challenges are to increase our funding levels, find a content developer and get to a steady production level. Short-term goals include publishing more lecture videos, developing course assignments and processing feedback from both experts and the wider AI safety community. Long-term goals include finishing the first course unit, developing a detailed course structure and making an accurate timeline.

Content

Milestones

RAISE was established in July 2017 as an initiative to improve the pipeline for AI safety research by creating an open online course. Since then, we built a steady fundament and published the first course material. Our milestones so far:

- **Formed reliable team**
  Toon Alfrink, Johannes Heidecke, Remmelt Ellen and Veerle de Goederen have committed to managing RAISE. Robert Miles has joined the team as our recorded lecturer.
- **Formalised mission statement**
  Our vision and mission are published in section 2 of this report.
- **Built study group**
  The online study group is dedicated to creating course content. Five to ten people attend each week. We have started with the topic of corrigibility. So far, we made summaries of nine research papers, wrote seven script drafts for video lectures, and prepared several presentations on more advanced topics in corrigibility.
- **Published prototype lesson**
  Robert Miles started making lecture videos, he finished a first video on moral uncertainty. The video was used to make a prototype lesson, consisting of short video clips interleaved with questions.

- **Designed logo and website**
  Jessica Mary Cooper designed our logo and website. We are thrilled by the results – thank you Jessica!
- **Received first funding**
  More on funding in section 3.

Vision and mission

To ensure that the creation of smarter-than-human artificial intelligence will have a positive impact, the world is in dire need of AI safety experts. But the path to becoming a capable researcher is complex and poorly defined. So far, few published papers have been compiled into a pedagogically-sound learning format – one that would enable aspirants to level up fast.

That's why we are building RAISE. Our mission is to unlock the capacity of aspiring AI safety researchers by creating an online course with high-quality content that's clear, distilled, community-centred, and open to everyone.

Our values

- **Accessibility** through material that is freely available online
- **Accountability** for our work, continually reporting progress and processing feedback to ensure well-tailored material.
- **Commitment to excellence** to bring material that is accurate, clear-cut, relevant and in-depth.
- **Teamwork** within our organisation, with our partners, and with the wider community.

Finances

RAISE currently has four donors, adding up to an income of €500 per month. So far, we have received €4714, of which we have spent €1560. Our current balance sits at approximately €3000.

Table 1 provides an overview of our estimated expenses, given we receive enough funding. Considering this expectation, our funding target is to raise about €4000 per month. It is most essential to fund a content developer and Robert, which together will cost around 1100 euros per month (financial compensation for two days per week). We do not expect sharp diminishing returns until well beyond €20.000 per month.

Table 1 Prospective funding. Amounts are given in euros per month.

	Lower bound	Upper bound
Content developer	640	3200
Robert	450	3200
Remmelt	200	1600
Software developer	640	3200
Toon	200	1600
Veerle	200	1600
Johannes	200	1600
Miscellaneous	100	500
Animation	700	3000
Video editing	625	2500
Total	3955	22000

Challenges and concerns

- **Content development**
  Currently, the course content is developed by volunteers. It appears challenging to make high-quality content at a steady rate, as most volunteers lack time and expertise.

  We want to resolve this by employing a skilled course content developer. The content developer will distil literature, develop explanations and design course assignments. The study group volunteers will provide the developer with support and feedback.

- **Production speed**
  Our production of course material so far has been slow. This is due to several reasons, most importantly the challenge of developing a framework for the course, the lack of a content developer and challenges that Robert ran into when making the first video.

  Over the next months, we want to work towards having a steady production level of course content. Our strategy is to find a content developer, and to support Robert as much as possible in making lectures. After three months, we aim to have clarity on the timeline of the course production.

- **Funding**
  Our main bottleneck at this point is funding. It is our priority to ensure financial compensation for a course content developer. In addition, we aim to provide financial compensation for our lecturer, a video editor/animator, a software developer, and the management team, in approximately that order.

  Our next steps are to approach several organisations and reach out to the AI safety community to ask for donations.

- **Product-market fit**
  Our target audience for RAISE are those who:

  - Already deem AI-alignment an important problem
  - Consider contributing to the field, potentially as researchers
  - Find it comparatively difficult to learn about AI safety by themselves considering how broad and ill-defined the field is

It is currently uncertain how much value RAISE can have. We do not yet have good estimates of how many people who are committed and gifted to do AI safety research will be reached by the course, and how useful the course will be for them.

Although we are mindful of the opportunity costs of funders and course contributors, we think there is a relatively low cost to RAISE in the case that it does not seamlessly match the needs of our target audience. On the upside, RAISE's positive impact could be substantial. At each step of content production, we will ask feedback from the AI safety community and adjust our progress accordingly. This feedback process will lead to better estimates of the value of an open online course on AI safety and allow us to optimise RAISE's positive impact. In addition, we will survey the added value of organised study groups and a wider learning community to explore if this should be part of our strategy.

Future

Based on our accomplishments so far and the challenges we are currently facing, we have set the following goals for the next three months:

- **Increase funding**
  In three months' time, we aim to have enough funding to compensate a content developer for at least half a year. Preferably, the inflow of donations will be steady, and we will have enough funding to also compensate our lecturer.
- **Employ content developer**
  We aim to find someone with the right skills who will boost our progress.
- **Publish videos**
  Our current estimate is that we can publish 2-5 more videos on corrigibility in the next few months.
- **Develop course assignments**
  The purpose of course assignments is to revise concepts, build confidence, obtain a more thorough understanding and stimulate creative solutions. Several types of assignments are needed to reach these different goals. In three months' time, we aim to have developed and tested several types of questions for the corrigibility unit, both revision questions and in-depth questions.
- **Process feedback**
  During the next months, we will ask feedback on published videos and assignments, both from the general AI safety community and from several experts. The feedback will be used to optimise the course content and teaching style.

Our future track consists of several major longer-term goals:

- **Finish first course unit**
  Finish the course unit on corrigibility, complete with lecture videos, accompanying text and course assignments.

- **Develop course structure**
  The structure scheme should give an overview of all course units, the main ideas covered in these units, their respective order, and the way course units relate to each other.

- **Set up detailed timeline**
  When we finished a first unit and developed a detailed course structure, we can make a better estimate of timescales. We will make a detailed timeline and set target dates for each unit. In addition, after tracking the reception of the first course unit we will evaluate how much money and skilled labour we should invest in future units.

The next status report will be published June 2018.