Decision-relevant uncertainty in AI safety

Owen Cotton-Barratt

Future of Humanity Institute, University of Oxford
Centre for Effective Altruism

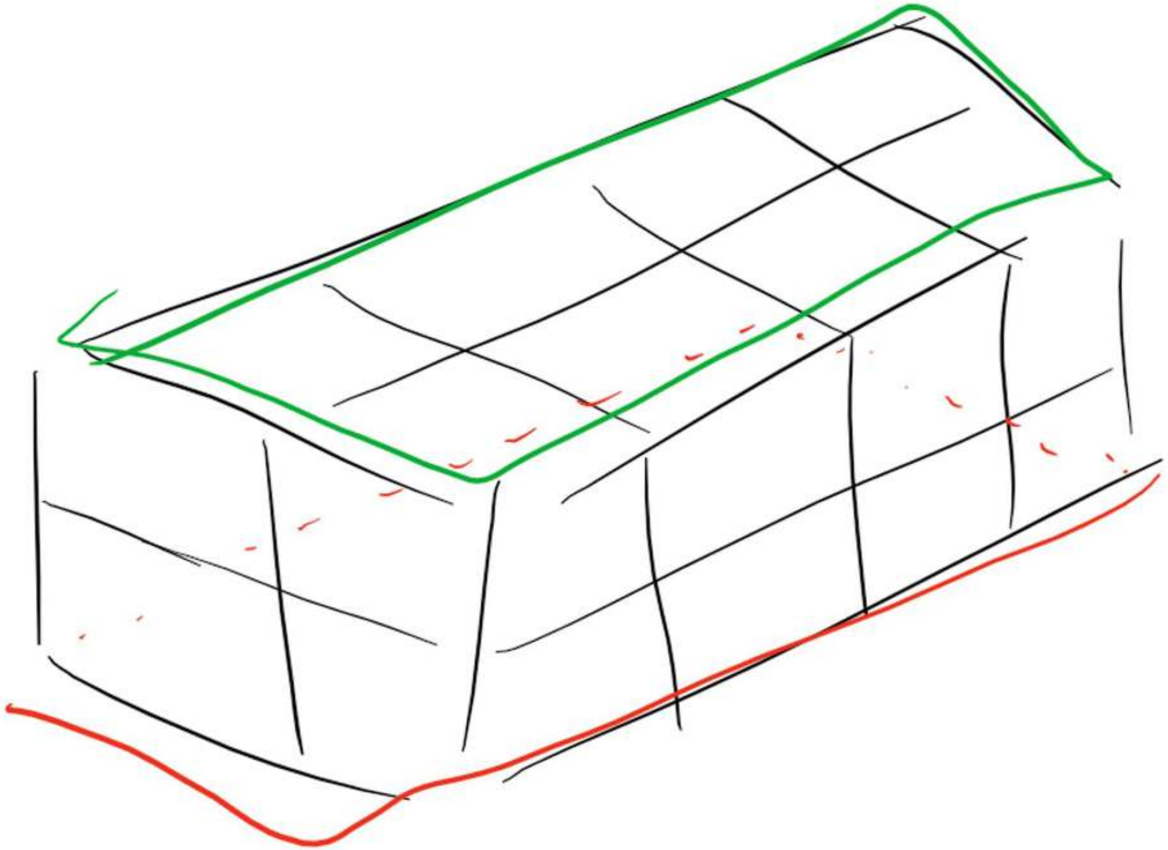
Central question: what to work on?

- What *portfolio* of work do we want?
- How is this affected by our uncertainty?
- This talk: focus on framing, not answers

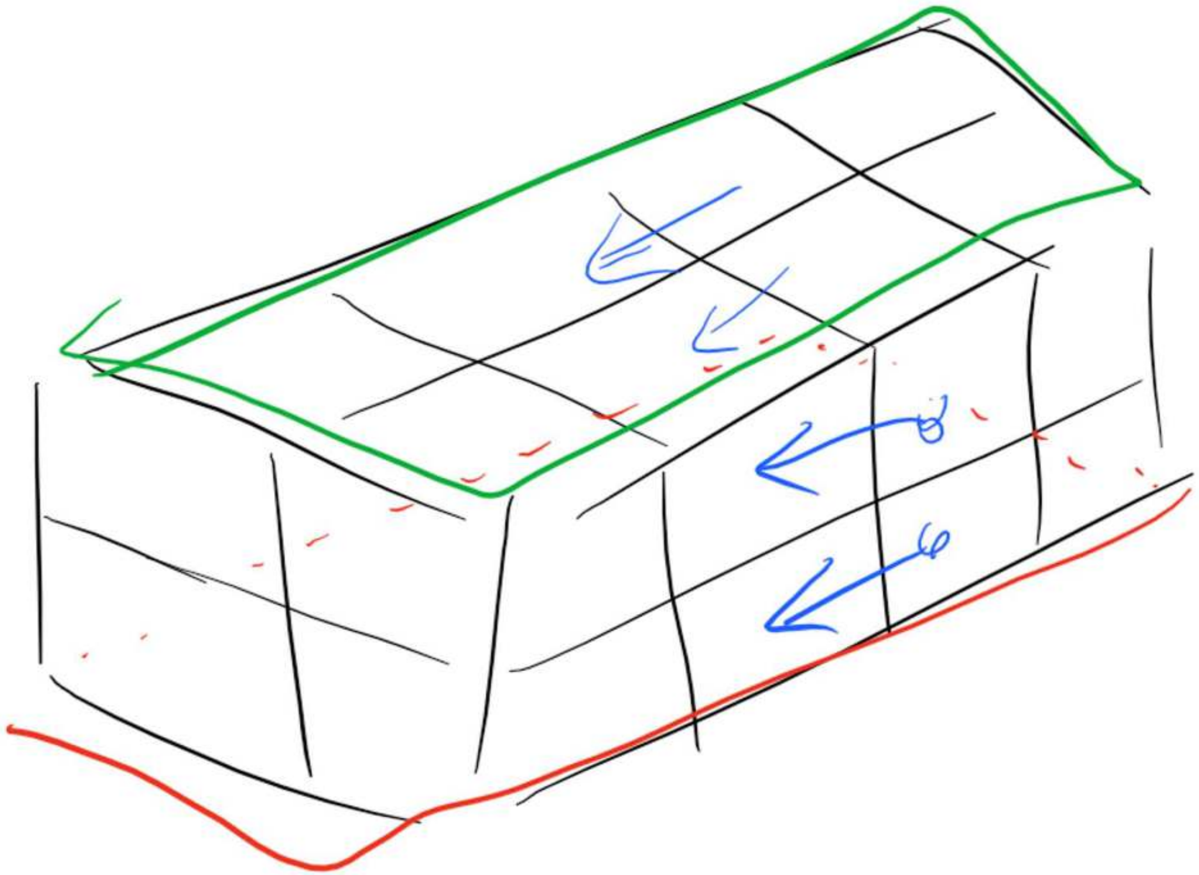
Some variables of uncertainty re. AI

- Timelines
  - When will we get AGI?
  - How fast/discontinuous will progress be?
  - How explosive could an intelligence explosion be?
- Development setting
  - How many major research groups will there be?
  - Will AGI be developed by industry/government/academia?
  - Will there be arms races?
  - How open will AI development be?
  - How far ahead of deployed capabilities will AI in development be?
- Technology
  - What technologies will the first AGI be based on?
  - How transparent will early AGI systems be?
  - Will there be game-changing implications of AI capabilities before AGI?
  - To what degree will AI systems remain ‘tool-like’?
- Safety work
  - What is the most likely cause of a catastrophe?
  - How much safety work will there be?
  - Which problems will be worked on?

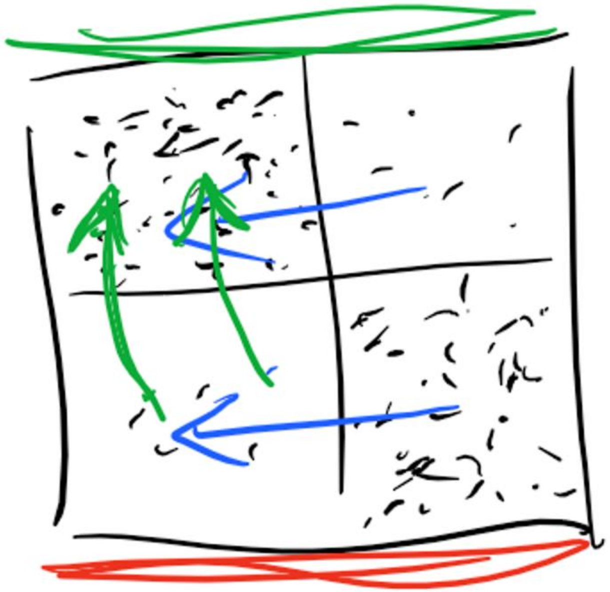
Uncertainty space



Acting in uncertainty space

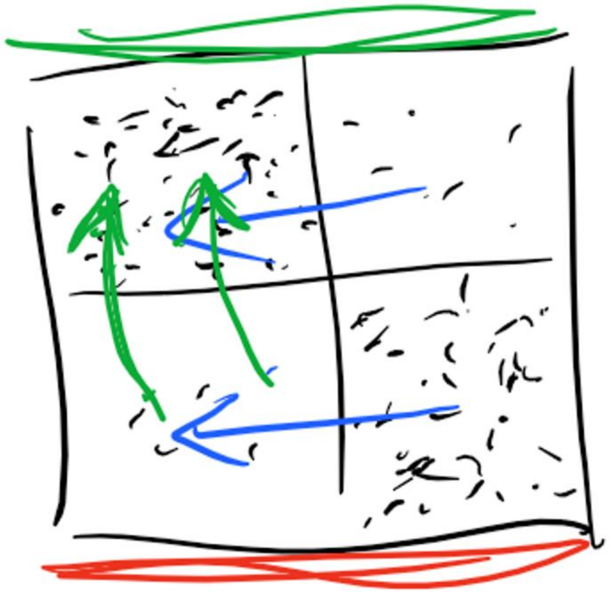


Variables to try to change

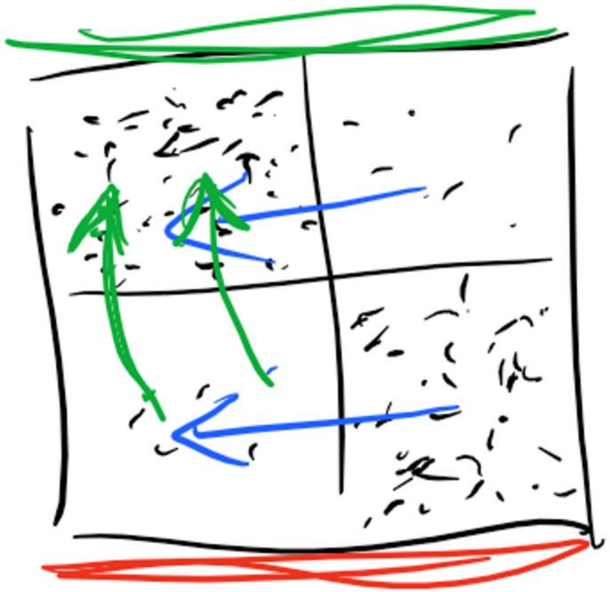


Variables to try to change

- We can’t target all variables



Variables to try to change

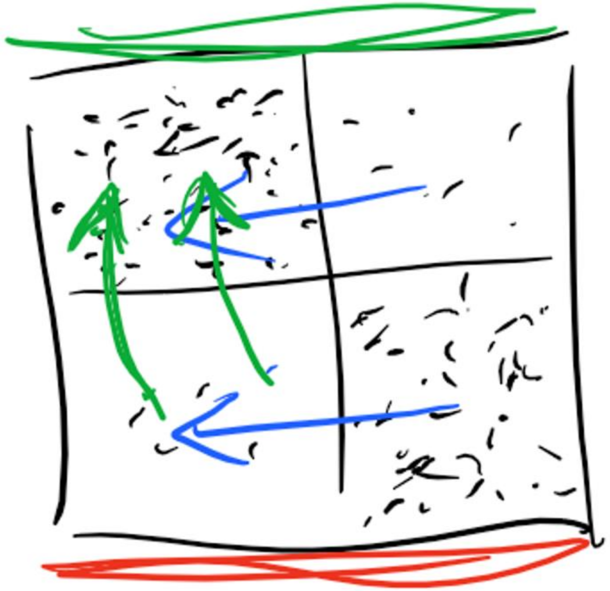


- We can't target all variables
- Desirable:
  - We can significantly affect the variable
  - The variable significantly affects probability of good outcomes
  - We know what direction is good

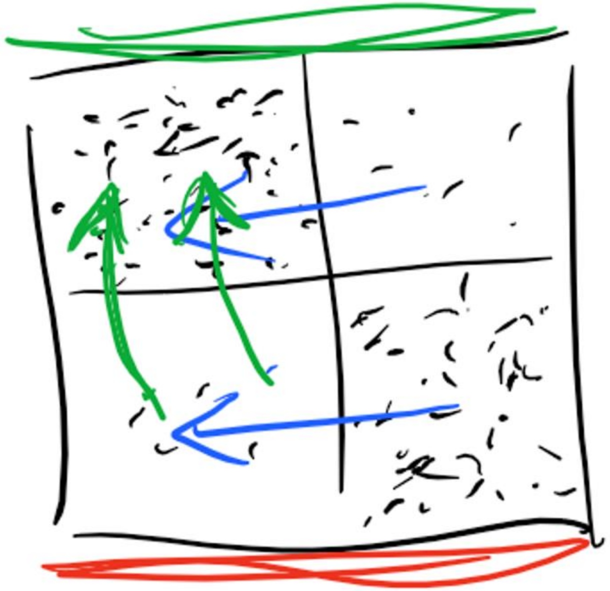
Variables to try to change

- We can't target all variables

- Desirable:
  - We can significantly affect the variable
  - The variable significantly affects probability of good outcomes
  - We know what direction is good



Variables to try to change



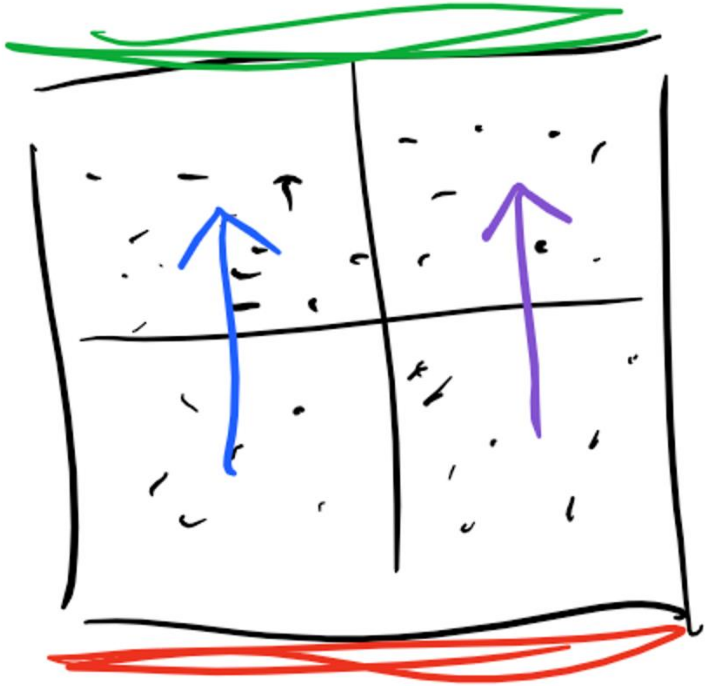
- We can't target all variables
- Desirable:
  - We can significantly affect the variable
  - The variable significantly affects probability of good outcomes
  - We know what direction is good
- Examples:
  - Good to target: amount of well-directed safety work
  - Bad to target: timelines

Other decision-relevant uncertainty

Do variables we're not trying to change still matter for our decisions?

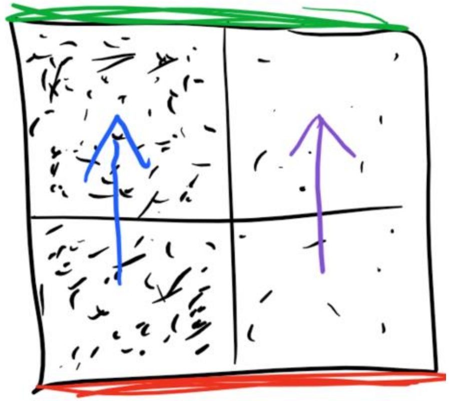
Yes, if we'd want to pursue different work conditional on different values of the variable.

e.g. 50-year vs 5-year timelines



Preferred work depends on …

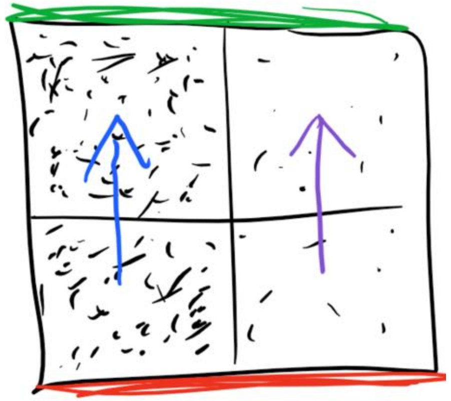
… likely outcomes



e.g. focus on ML

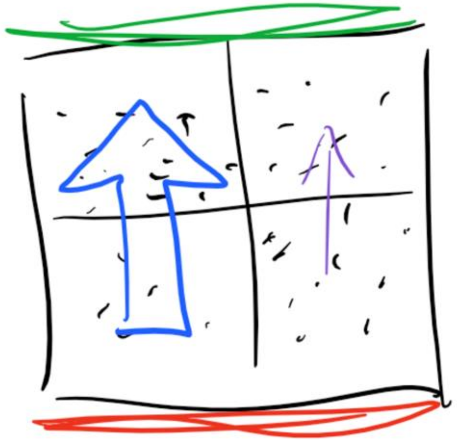
Preferred work depends on …

… likely outcomes



e.g. focus on ML

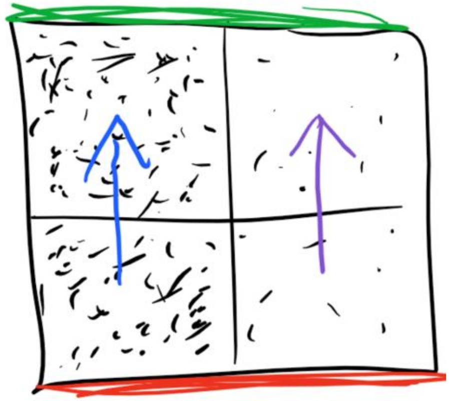
… leverage



e.g. focus on unexpectedly short timelines

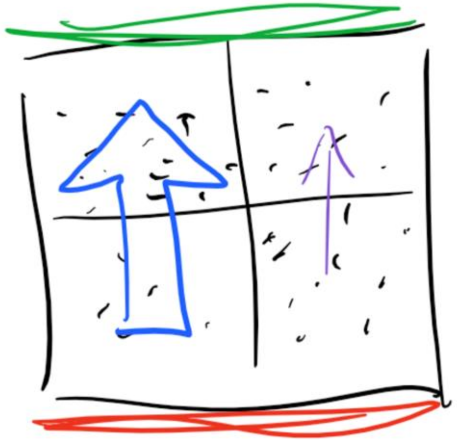
Preferred work depends on …

… likely outcomes



e.g. focus on ML

… leverage



e.g. focus on unexpectedly short timelines

Ratio of value of work = (odds ratio) x (leverage ratio)