Combinatorial Questions for AI Safety

Laurent Orseau
with Pedro Ortega, Victoria Krakovna, Jan Leike, Shane Legg
PgM: Andrew Lefrancq

AI Safety

- **AI Safety: Reduce the risks of something going wrong**

- **Known unknowns** (uncertainty)
  - When a problem is revealed, many people can work on it
  - Reduce the risk of a known problem

- **Unknown unknowns** (incomplete knowledge)
  - Problems yet to be discovered
  - Due to bounded rationality

How to make sure we don't miss important problems?

Combinatorial Questions

- Identify a single question
  - Ex: What if the distribution of the examples can change during training?

- Identify categories: actors, actions, objects, places, times, etc.
  - Ex: Objects = {examples}
    Times = {training}

- For each category, think of other possible values
  - Ex: Objects = {examples, class-labels, agent-answers}
    Times = {training, testing, use-time}

Combinatorial questions

- Ex: What problems can occur if…

$$\text{the distribution of the } \begin{bmatrix}\text{examples}\\ \text{class-labels}\\ \text{agent-answers}\end{bmatrix} \text{ changes during } \begin{bmatrix}\text{testing}\\ \text{training}\\ \text{use-time}\end{bmatrix} ?$$

- Examples, training
  - Catastrophic forgetting
- Class labels, training
  - Tracking
- Examples, testing
  - Extrapolation/generalization issues
- Agent answers, use-time
  - Volkswagen problem
- …

Combinatorial Questions

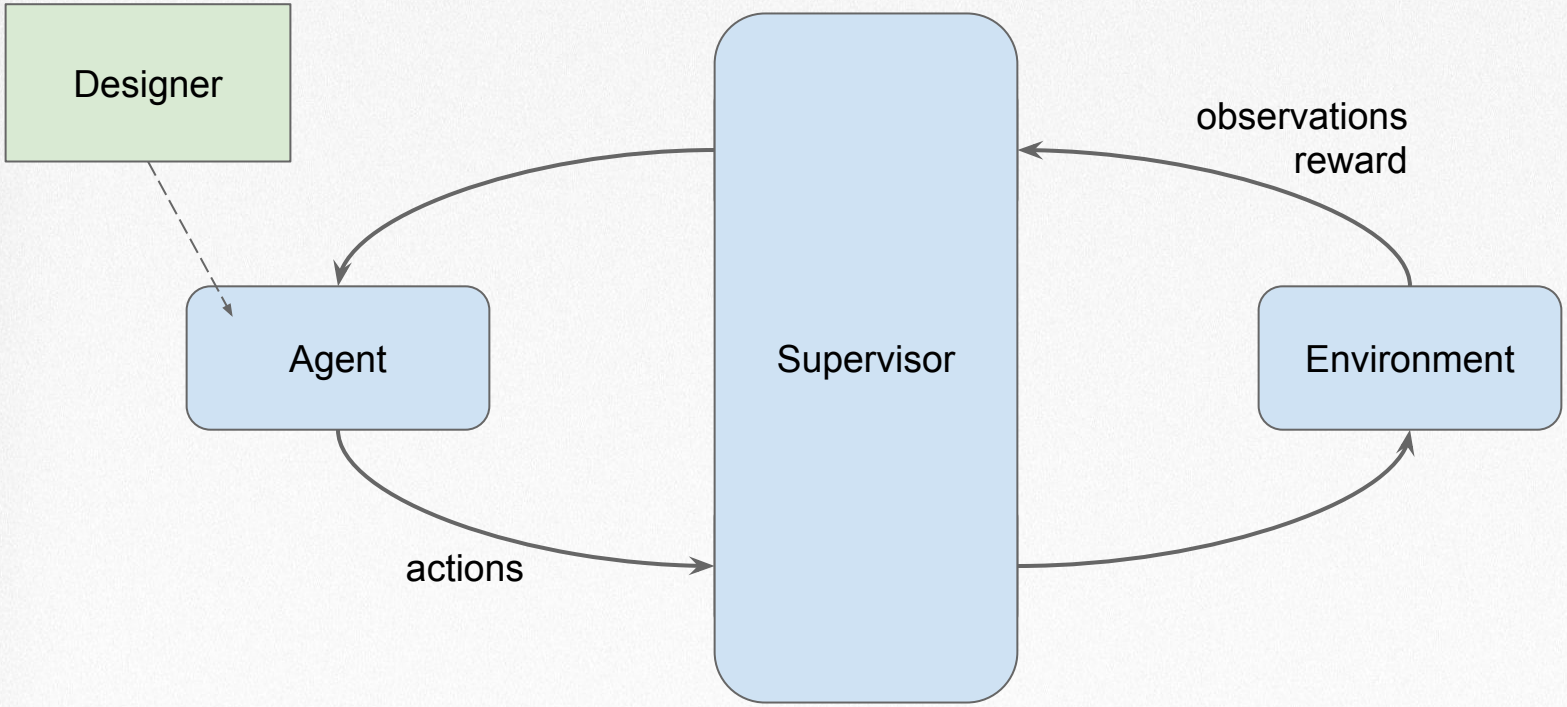
Systematic way to generalize one discovered problem to many potential ones

Bring chunks of **unknown unknowns** to **known unknowns**

What to do with a combinatorial question?

- Not all questions make sense
  - But use your imagination!
- Not all questions are equally important
  - Maybe try to reduce the set?
- If #combinations small, try them all
- If too large,
  - Sample from the larger set
    - Something interesting may come up!
  - Build 1 or several reduced questions

Agent Environment Framework



Modification questions

What problems can arise if…

(800 questions)

Modification questions: Some existing work

- What if…
  - The environment can read the policy of the agent?
    - Many decision theory problems (Newcomb, etc.)
  - The environment can {move, duplicate} the memory of the agent?
    - Teleportation and identity [Orseau&Ring, 2014]
  - The supervisor can modify the reward function of the agent?
    - IRL [Ng&Russell 2000], Corrigibility [Soares et al., 2015]
  - The supervisor can modify the actions of the agent?
    - Safe interruptibility [Orseau&Armstrong, 2016]
  - The agent can modify the reward function/policy of the agent?
    - Self-modification [Orseau&Ring, 2011]
  - The agent can modify the observations of the agent?
    - Delusion Box problem [Ring&Orseau, 2011]
  - The environment can change the reward function of the agent?
    - Jekyll&Hyde problem (unpublished)
  - …

Modification questions: New questions?

- What if…
  - The agent can delay the observations of the supervisor ?
  - The agent can modify the actions of the supervisor?
  - The agent can delay the observations of the agent?
  - The agent modifies the actions of the agent?
    - E.g., modifies its set of available actions
  - …

Trust Questions

- **(AI) safety is all about trust**
  - Trust in the future behaviour of the agent
  - Trust that the goals/intentions of the designers are

- One question:

How can the designer gain more trust in the agent?

- Actors = {designer, agent, user, press, people, government, …}

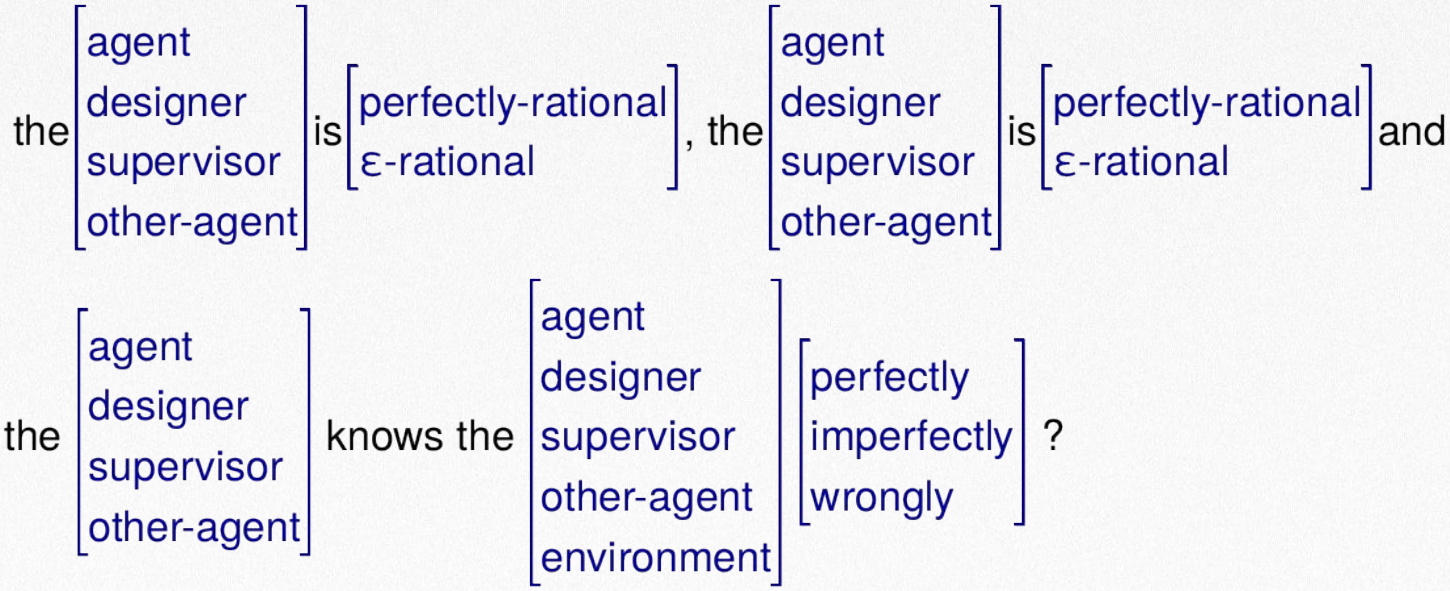
Trust questions

How can the [agent / designer / supervisor / users / people / client / government / press / company / company-competitors] gain more trust in the [agent / designer / supervisor / users / people / client / government / press / company / company-competitors] ?

(100 questions)

Uncertainty Questions

What problems can arise if…



(3840 questions)

Uncertainty questions: Samples

- What can go wrong when…
  - The designer is ε-rational, the agent is ε-rational and the designer knows the agent imperfectly?
  - The agent is rational, the designer is ε-rational, and the agent knows the designer perfectly
  - The agent is rational, the designer is ε-rational, and the designer knows the environment imperfectly?
  - The other-agent is rational, the agent is ε-rational, the other-agent knows the agent perfectly

Conclusion

- Combinatorial questions help turning unknown unknowns into known unknowns

- What optimal cardinality?
  - Need to generalize to other cases
  - Avoid including too many meaningless questions
  - Learn
    - what is likely to be interested
    - what is least covered?

- What other interesting combinatorial questions?