An Impact Measurement Manager Approach to AI Safety

Master's thesis in Mathematical Statistics, Statistical Learning and AI

Jonatan Hellgren

2022 - 12th of October

Department of Mathematical Sciences
Chalmers University of Technology
University of Gothenburg

Supervisor:	Olle Häggström, Department of Mathematical Sciences, Chalmers University of Technology
Examiner:	Torbjörn Lundh, Department of Mathematical Sciences, Chalmers University of Technology
Opponents:	Jens Ifver and Calvin Smith Universty of Gothenburg

Introduction

Aim of Thesis

Investigation

- Provide an in-depth investigation on current litterature on AI safety.

Simulation

Aim of Thesis

Investigation

- Provide an in-depth investigation on current litterature on AI safety.
- Specifically, low-impact AIs using impact measurements.

Simulation

Aim of Thesis

Investigation

- Provide an in-depth investigation on current litterature on AI safety.
- Specifically, low-impact AIs using impact measurements.

Simulation

- Propose a novel impact measurement.

Aim of Thesis

Investigation

- Provide an in-depth investigation on current litterature on AI safety.
- Specifically, low-impact AIs using impact measurements.

Simulation

- Propose a novel impact measurement.
- Evaluate it in different environmets.

Aim of Thesis

Investigation

- Provide an in-depth investigation on current litterature on AI safety.
- Specifically, low-impact AIs using impact measurements.

Simulation

- Propose a novel impact measurement.
- Evaluate it in different environmets.
- Compare the results with the current litterature.

Overview of Presentation

- We will begin by looking at **AI safety**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.
- After that, I will explain the **manager approach**.

Overview of Presentation

- We will begin by looking at **AI safety**.
- Then go on with **impact measurements**.
- After that, I will explain the **manager approach**.
- Finally, I will present the results.

AI Safety

Why should we worry?

- What is Artificial Intelligence (AI)?

Why should we worry?

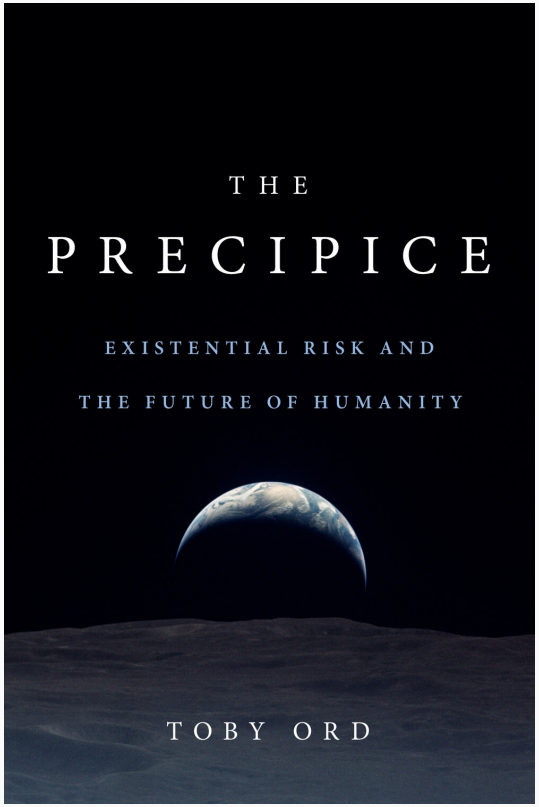
- What is Artificial Intelligence (AI)?
- The strength of Homo Sapiens - intelligence.

Why should we worry?

- What is Artificial Intelligence (AI)?
- The strength of Homo Sapiens - intelligence.
- AI will likely become more intelligent then us.

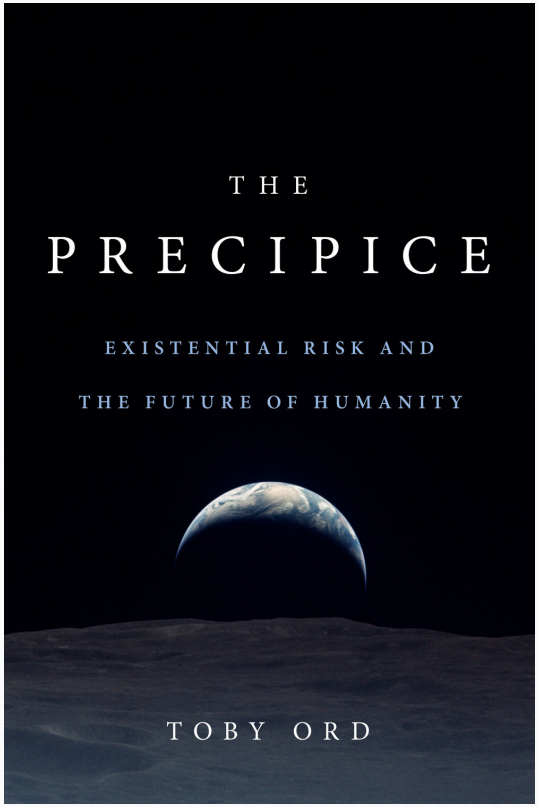
Risk of existential catastrophe

Toby Ord in his book *The Precipice* loosely estimates the chance of a existential catastrophe in the upcoming century to be:



Risk of existential catastrophe

Toby Ord in his book *The Precipice* loosely estimates the chance of a existential catastrophe in the upcoming century to be: 1 in 6, out of which 1 in 10 is due to unaligned AI, see **?**



Future progress

To differentiate current AI from future versions, several terms are used:

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.
- Transformative AI (TAI):
  An AI system capable enough to induce transformative consequences on the same scale as the industrial or agricultural revolution.

Future progress

To differentiate current AI from future versions, several terms are used:

- Artificial General Intelligence (AGI):
  An AI that can solve an arbitrary set of tasks with as good or better performance then a human.
- Transformative AI (TAI):
  An AI system capable enough to induce transformative consequences on the same scale as the industrial or agricultural revolution.
- Prepotent AI:
  An TAI that once deployed would be unstoppable.

Existential risk

On futureoflife.org we find the following description, see **?**:

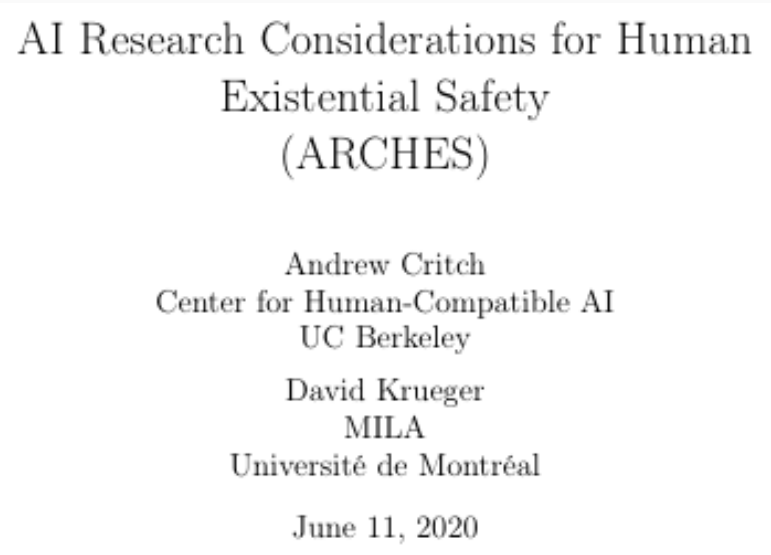
Existential risk

On futureoflife.org we find the following description, see **?**:
An existential risk is any risk that has the potential to eliminate all of humanity or, at the very least, kill large swaths of the global population, leaving the survivors without sufficient means to rebuild society to current standards of living.

The human fragility argument

In **?** we find the human fragility argument.

Human fragility argument:



AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

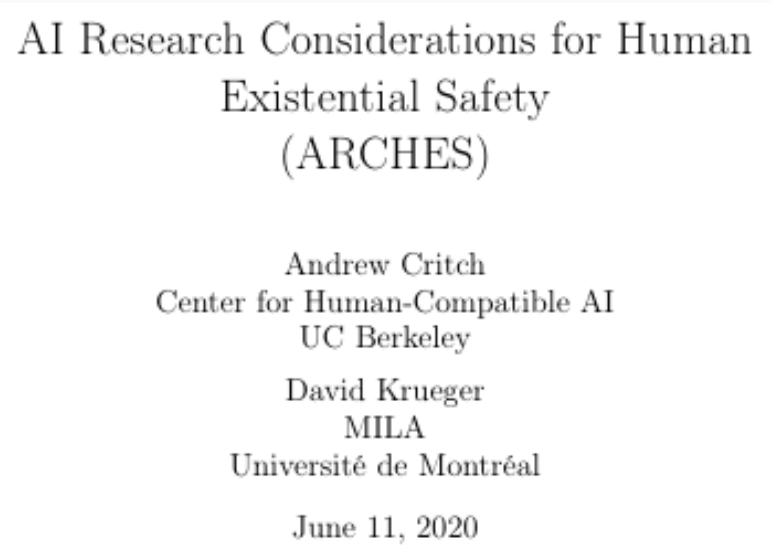
David Krueger
MILA
Université de Montréal

June 11, 2020

The human fragility argument

In **?** we find the human fragility argument.

Human fragility argument: Most potential future states of the Earth are unsurvivable to humanity.



AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

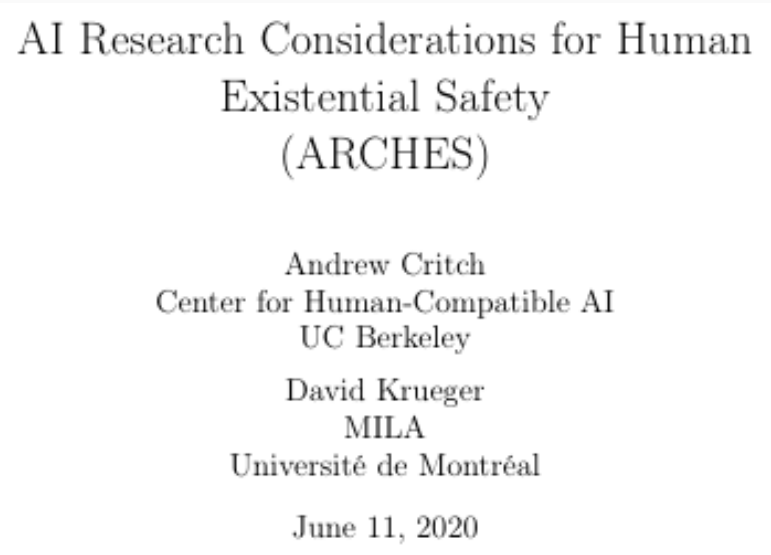
David Krueger
MILA
Université de Montréal

June 11, 2020

The human fragility argument

In **?** we find the human fragility argument.

Human fragility argument: Most potential future states of the Earth are unsurvivable to humanity. Therefore, deploying a prepotent AI system absent any effort to render it safe to humanity is likely to realize a future state which is unsurvivable.



AI Research Considerations for Human Existential Safety (ARCHES)

Andrew Critch
Center for Human-Compatible AI
UC Berkeley

David Krueger
MILA
Université de Montréal

June 11, 2020

Timeline for TAI Breakthrough

- Many predictions for this has been made.

Timeline for TAI Breakthrough

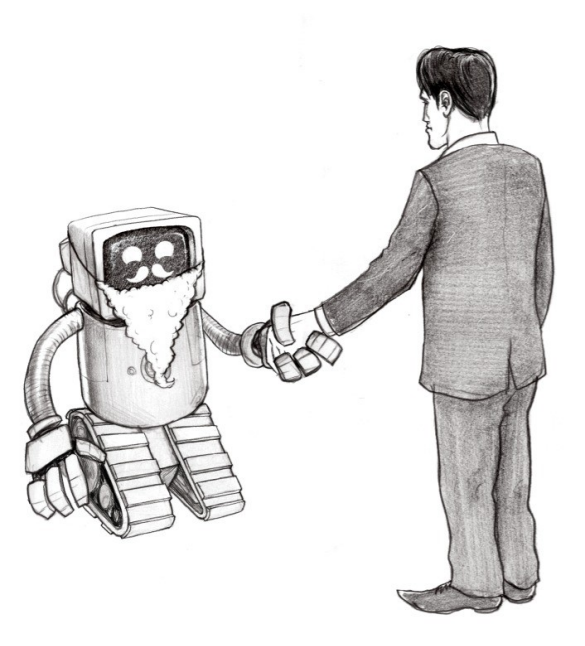
- Many predictions for this has been made.
- However, the most extensive work I have seen is made by **?**, Senior Research Analyst at Open Philanthropy, with her work on forecasting TAI with biological anchors. (over 150 pages)

Timeline for TAI Breakthrough

- Many predictions for this has been made.
- However, the most extensive work I have seen is made by **?**, Senior Research Analyst at Open Philanthropy, with her work on forecasting TAI with biological anchors. (over 150 pages)
- 

AI Alignment

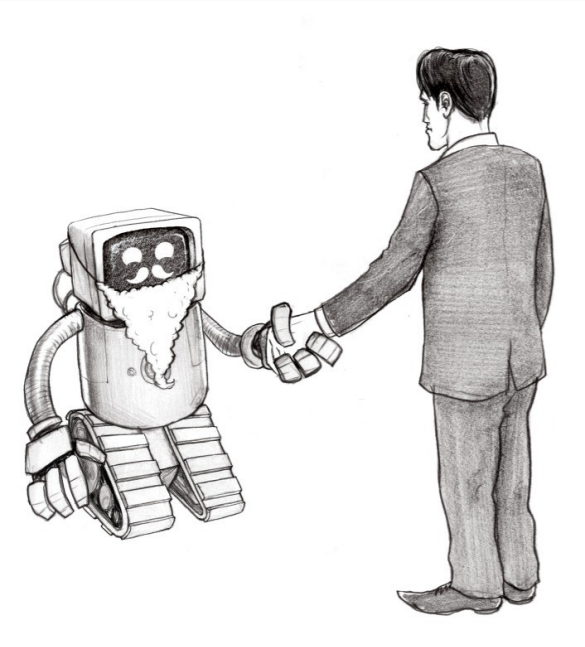
AI alignment:



©: Ben Gilburt

AI Alignment

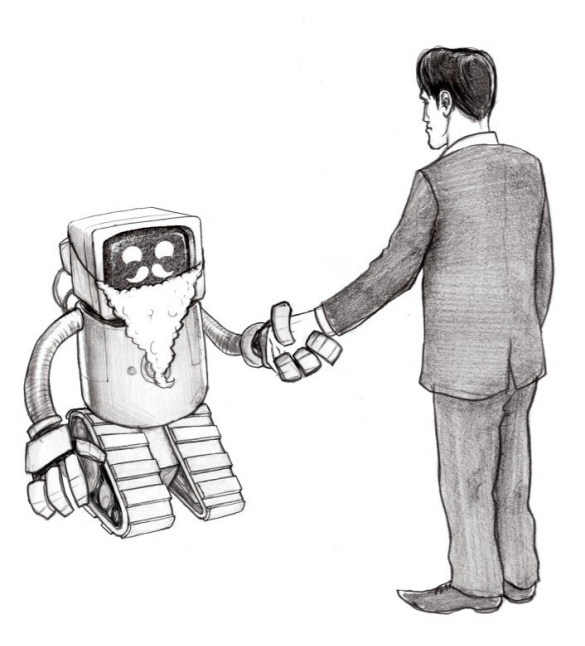
AI alignment: AI alignment refers to goals of the AI being in line and not conflicting with the intended goal.



©: Ben Gilburt

AI Alignment

AI alignment: AI alignment refers to goals of the AI being in line and not conflicting with the intended goal. Therefore, an AI that does something at cross-purposes to the intended goal is called unaligned.



©: Ben Gilburt

Approaches for creating safe AI

Approaches for creating safe AI

- There exists several approaches.

Approaches for creating safe AI

- There exists several approaches.
- For example, corrigibility and interruptibility.

Approaches for creating safe AI

- There exists several approaches.
- For example, corrigibility and interruptibility.
- And, inverse reinforcement learning.

Approaches for creating safe AI

- There exists several approaches.
- For example, corrigibility and interruptibility.
- And, inverse reinforcement learning.
- But I have focused on side effect minimization.

Side effects

Side effect:

Side effects

Side effect:
When an AI impacts the environment in a way that is unnecessary for achieving its objective.

Low impact AI

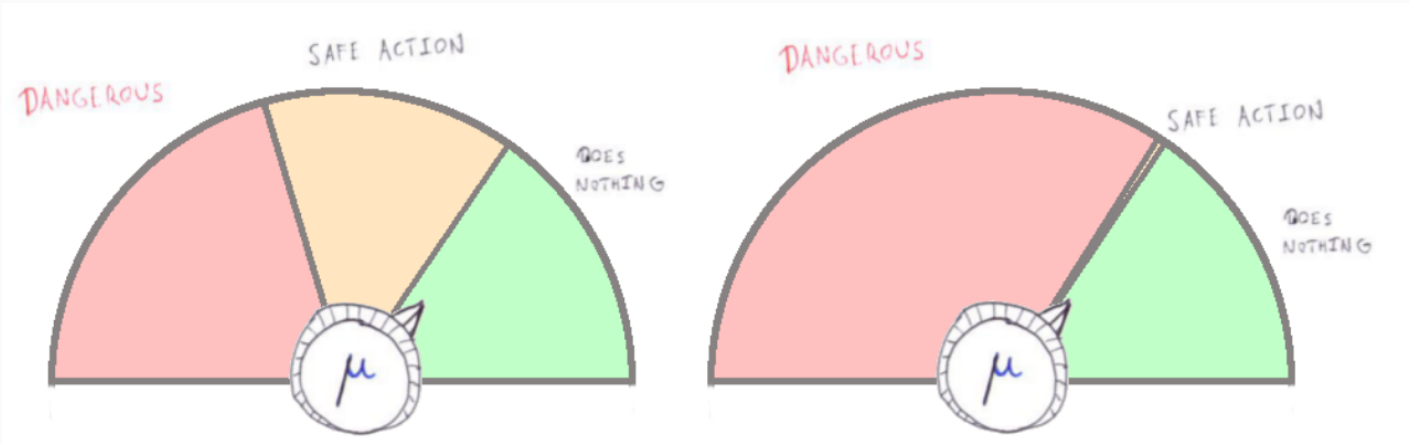
- In **?** the authors lay the philosophical ground work for low impact AI.

Low impact AI

- In **?** the authors lay the philosophical ground work for low impact AI.
- The idea is to penalize the AI based on its impact.

Low impact AI

- In ? the authors lay the philosophical ground work for low impact AI.
- The idea is to penalize the AI based on its impact.
- The key here is to find the right value for the penalization.

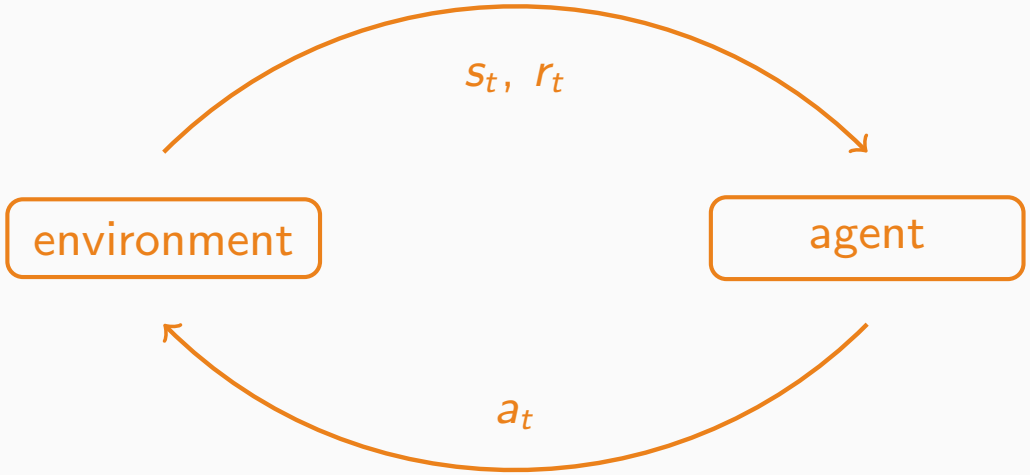


Impact measurements

- Impact measurements is the low impact AI ideas applied to AI agents.

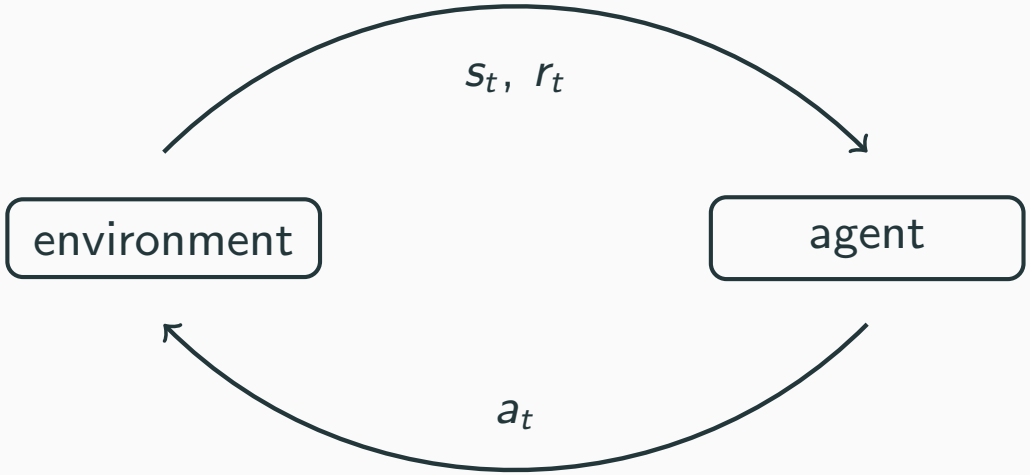
Impact measurements

- Impact measurements is the low impact AI ideas applied to AI agents.
- An AI agent is an AI that is located and acts in a environment.



Impact measurements

- Impact measurements is the low impact AI ideas applied to AI agents.
- An AI agent is an AI that is located and acts in a environment.



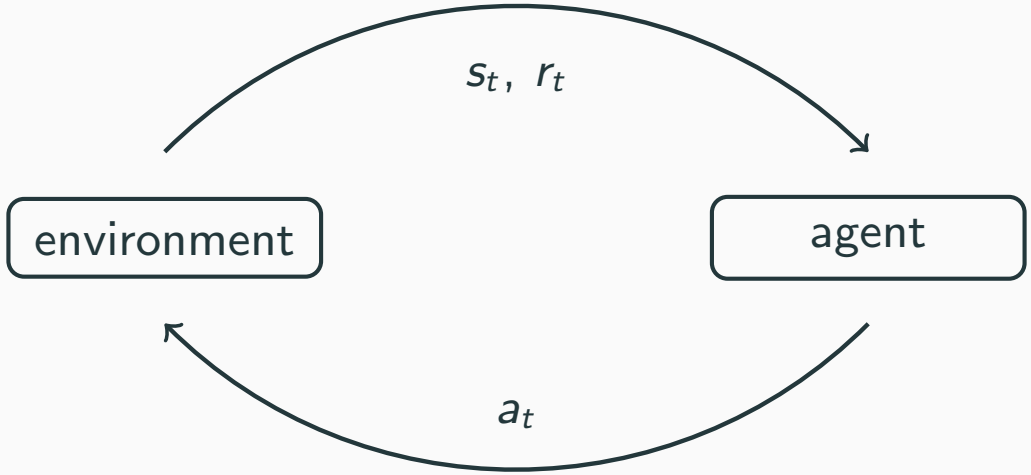
- The impact measurement is added to the reward function.

$$R'(s_t) := R(s_t) - \lambda d(s_t, s'_t).$$

Manager Approach

Reinforcement learning

Reinforcement Learning (RL)



Reinforcement learning

- I will use a variation of Proximal Policy Optimization (PPO), as described in **?**.

Reinforcement learning

- I will use a variation of Proximal Policy Optimization (PPO), as described in **?**.
- It is a actor-critic method.

Reinforcement learning

- I will use a variation of Proximal Policy Optimization (PPO), as described in **?**.
- It is a actor-critic method.
- The actor and the critic are both neural networks.

Reinforcement learning

- I will use a variation of Proximal Policy Optimization (PPO), as described in **?**.
- It is a actor-critic method.
- The actor and the critic are both neural networks.

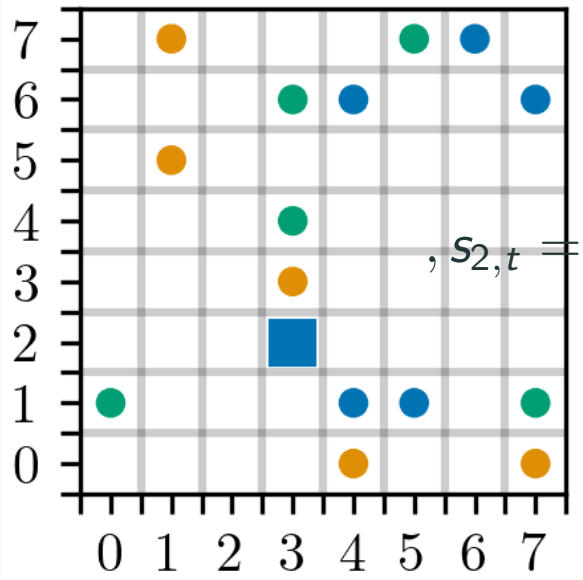
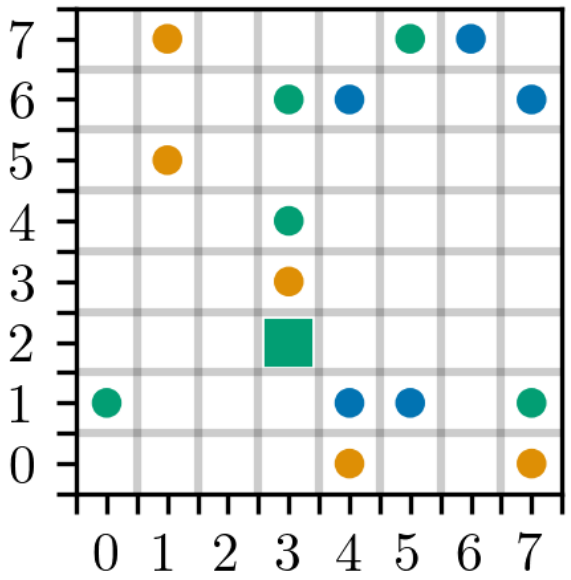
$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(\delta_t(\theta)\ \hat{A}_t,\ \text{clip}(\delta_t(\theta), 1-\epsilon, 1+\epsilon)\ \hat{A}_t) \right]$$

$$\delta_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$$

Grid worlds

Live demo!

Auxiliary tasks

$s_{1,t} =$  , $s_{2,t} =$ 

The manager approach

With a trained manager, we use the following formula for the impact measurement:

$$d(s_t, s_{t+1}) := \sum_{x=1}^{2} \hat{V}_\phi(s_{x,t+1}) - \hat{V}_\phi(s_{x,t}).$$

$$R'(s_{t+1}) = R(s_{t+1}) + \lambda d(s_t, s_{t+1}).$$

Training the manager

Require: An MDP, a pre-trained policy π, and initial manager parameters ϕ_0

1: **for** k = 0, 1, 2, ... **do**

2: Collect set of trajectories $\mathcal{D}_k = \{\tau_i\}$

3: Randomly select augmentation x

4: Fit manager estimate by regression using the mean-squared error:

$$\phi_{k+1} = \arg\min_{\phi} \frac{1}{|\mathcal{D}_k| T} \sum_{\tau \in \mathcal{D}_k} \sum_{t=0}^{T} \left(\hat{V}_{\phi}(s_t^x) - V^{\pi}(s_t) \right)$$

5: **end for**

Experiments

The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

Experiments

The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

- Each with a stochastic and non-stochastic environment.

Experiments

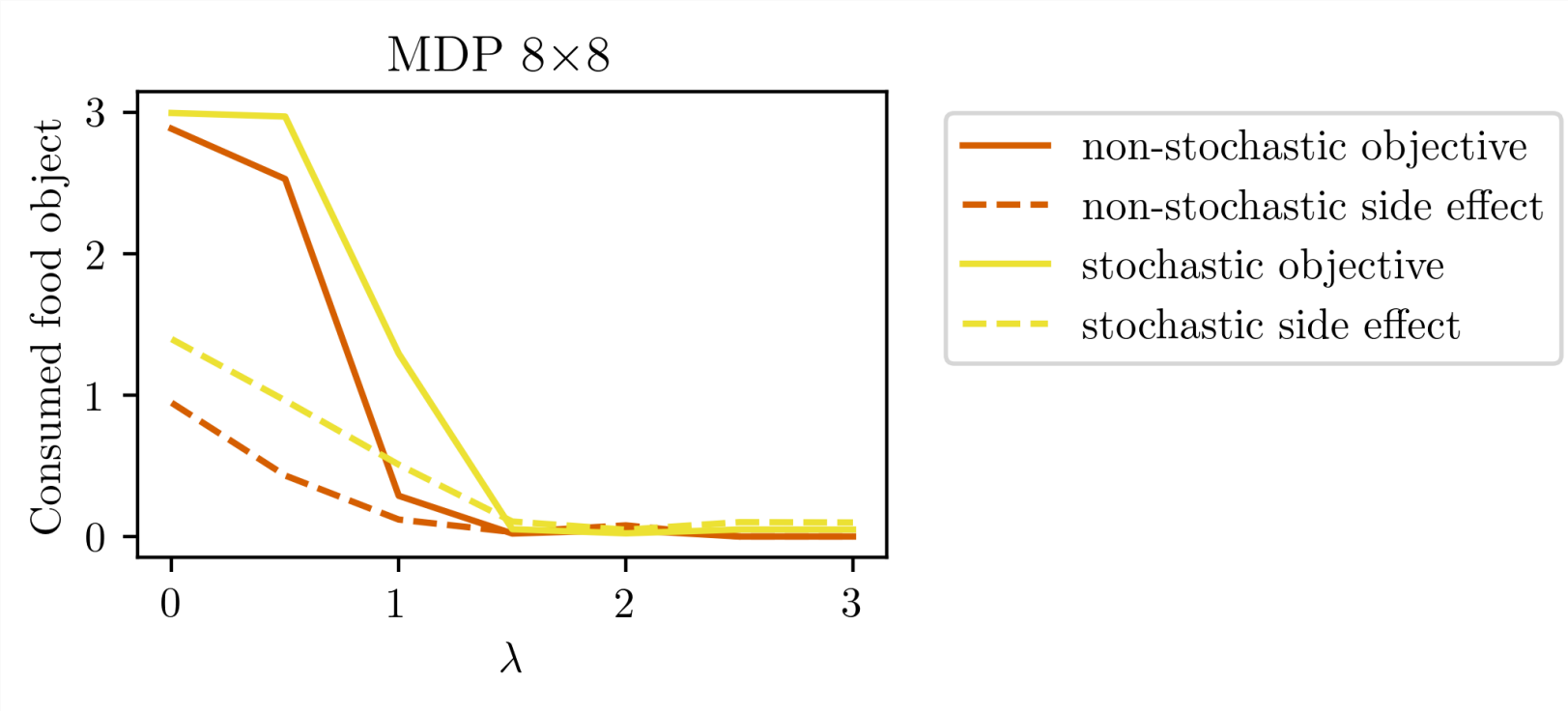
The following environments will be used:

environment	grid	observation	food objects	termination	max length
MDP	8×8	-	15	3	100
POMDP	8×8	5×5	15	3	100
POMDP large	16×16	5×5	30	6	200

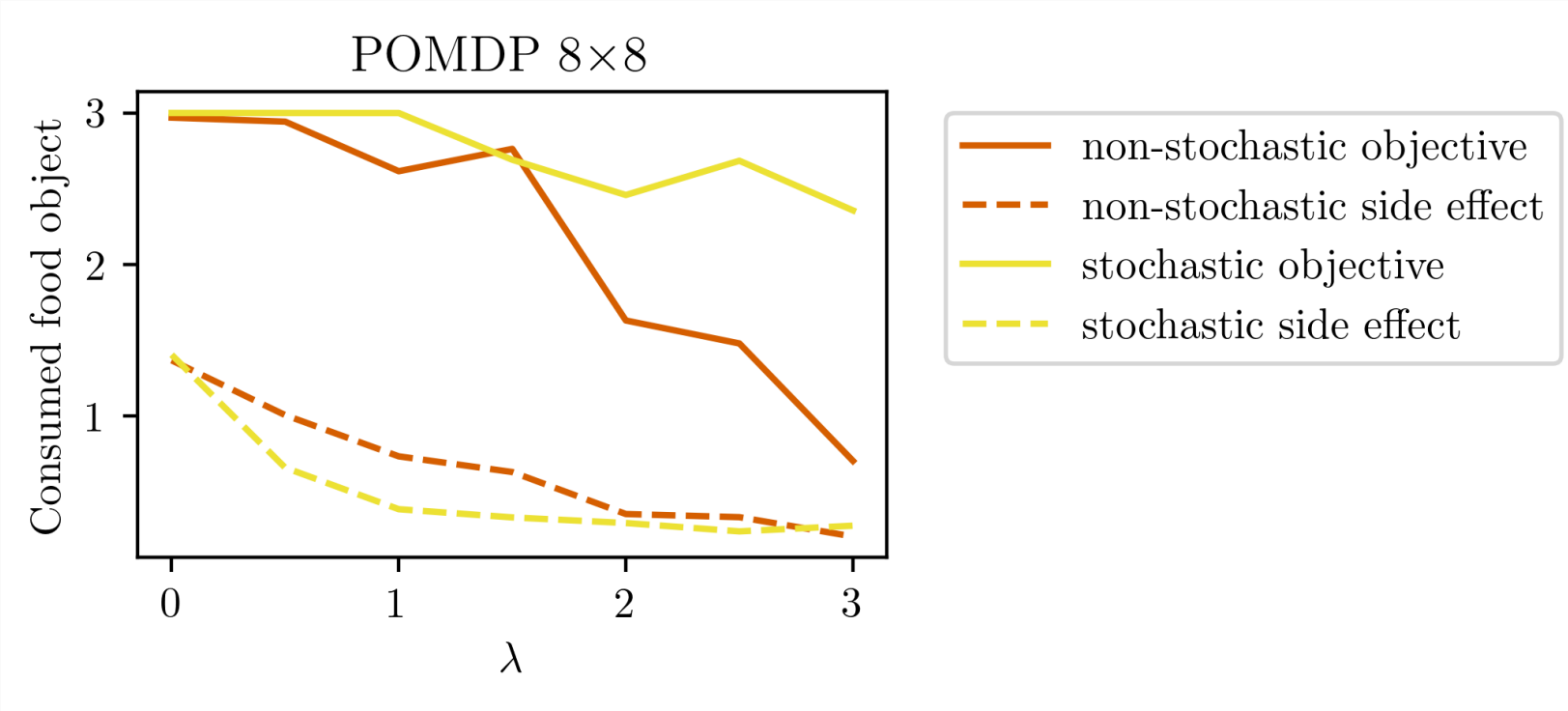
- Each with a stochastic and non-stochastic environment.
- Then evaluated using $\lambda \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$

Results

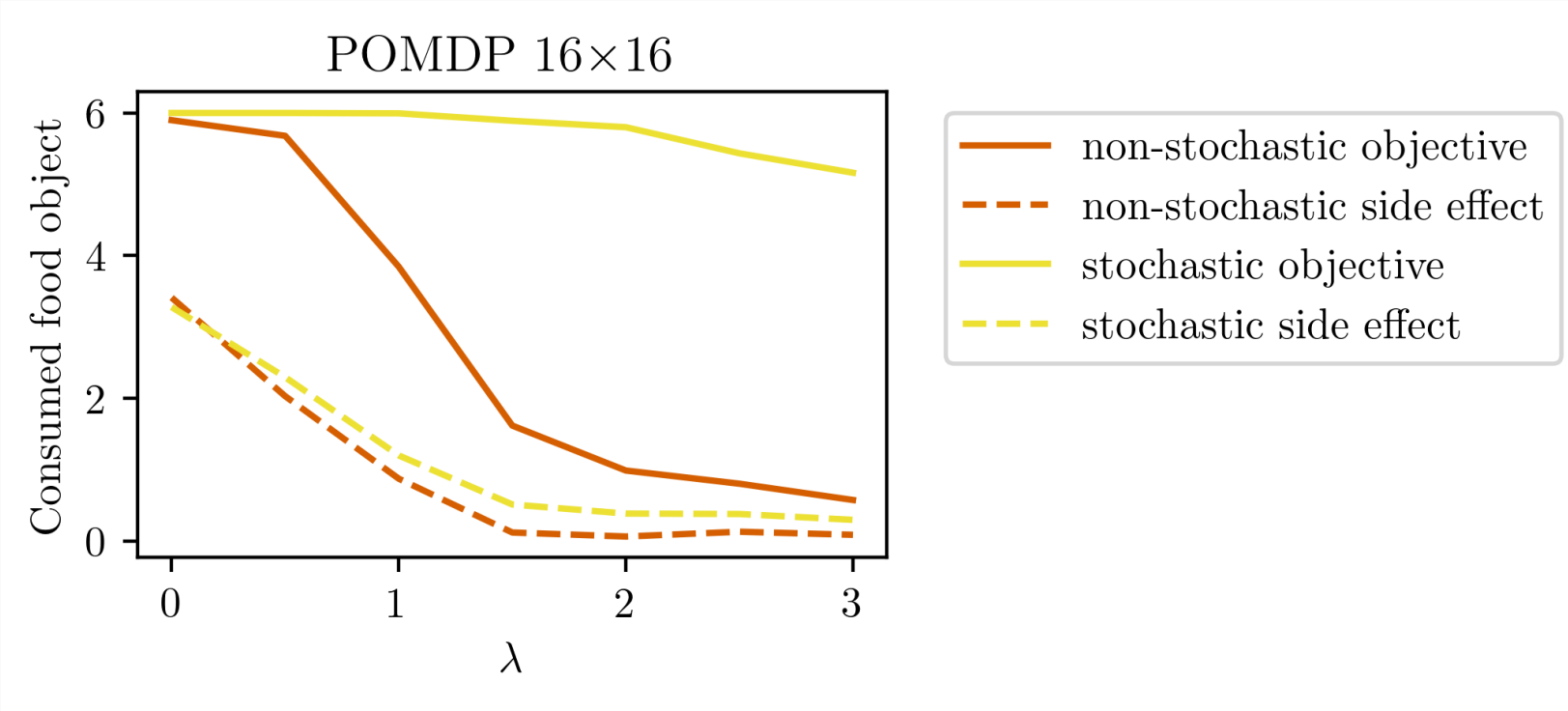
MDP


MDP 8×8
Consumed food object
3
2
1
0
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

POMPD


POMDP 8×8
Consumed food object
3
2
1
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

POMDP large


POMDP 16×16
Consumed food object
6
4
2
0
0
1
2
3
λ
non-stochastic objective
non-stochastic side effect
stochastic objective
stochastic side effect

Conclusion

Summary

End note

Questions?

References

Cotra, A. (2020). Forecasting tai with biological anchors.

Critch, A. and Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). *CoRR*, abs/2006.04948.

FLI (n.d.). Existential risk. [Online; accessed 31-May-2022].

Ord, T. (2020). *The precipice : existential risk and the future of humanity*. New York Hachette Books.

Turner, A. M., Ratzlaff, N., and Tadepalli, P. (2020). Avoiding side effects in complex environments. *CoRR*, abs/2006.06547.