Using Machine Learning to Address AI Risk

Jessica Taylor Eliezer Yudkowsky
Patrick LaVictoire Andrew Critch
{jessica,eliezer,patrick,critch}@intelligence.org

Machine Intelligence Research Institute

August 6, 2016

Outline of the workshop

- Talk ($\sim$ 45 mins)
  - Goal of this research agenda
  - 6 potential problems with highly capable AI systems
  - More technical depth about one research question
  - Other research agendas
  - Conclusion
- Questions, comments, discussion ($\sim$ 15 mins)

Goal of this research agenda

Assumptions behind this agenda

Goal statement: Know how to train smarter-than-human AI systems to perform large, useful tasks in the world.

Assumptions:

- Future AI systems might be substantially similar to present-day machine learning systems.
- AGI might be developed soon (in the next 20 years), and it is useful to focus on these short timelines.
- It's useful and tractable to research how to build a task-directed AI.

More details on task-directed AI

(Similar to "genie" in Bostrom's typology)

A *task* is a semi-concrete objective in the world.

- Build a million houses
- Cure cancer

Not:

- Learn human values and do things humans would consider good upon sufficient reflection

Hope: task-directed AI is sufficient to prevent global catastrophic risks

More details on task-directed AI

Task-directed AI will use moderate human assistance to clarify the goal and to evaluate or carry out plans.

Ideally, task-directed AI should not require much more computing resources than competing systems.

Modeling future AI systems

Future AI systems will use new algorithms, new data, and new hardware. How do we model it?

Modeling future AI systems

Future AI systems will use new algorithms, new data, and new hardware. How do we model it?

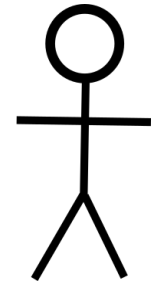
General approach: look at current AI systems and imagine more powerful versions of them.

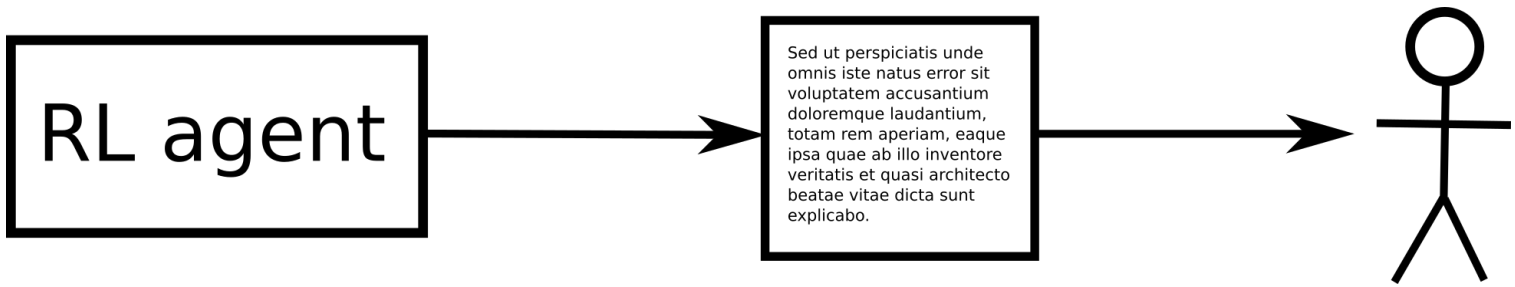
6 potential problems with highly capable AI systems

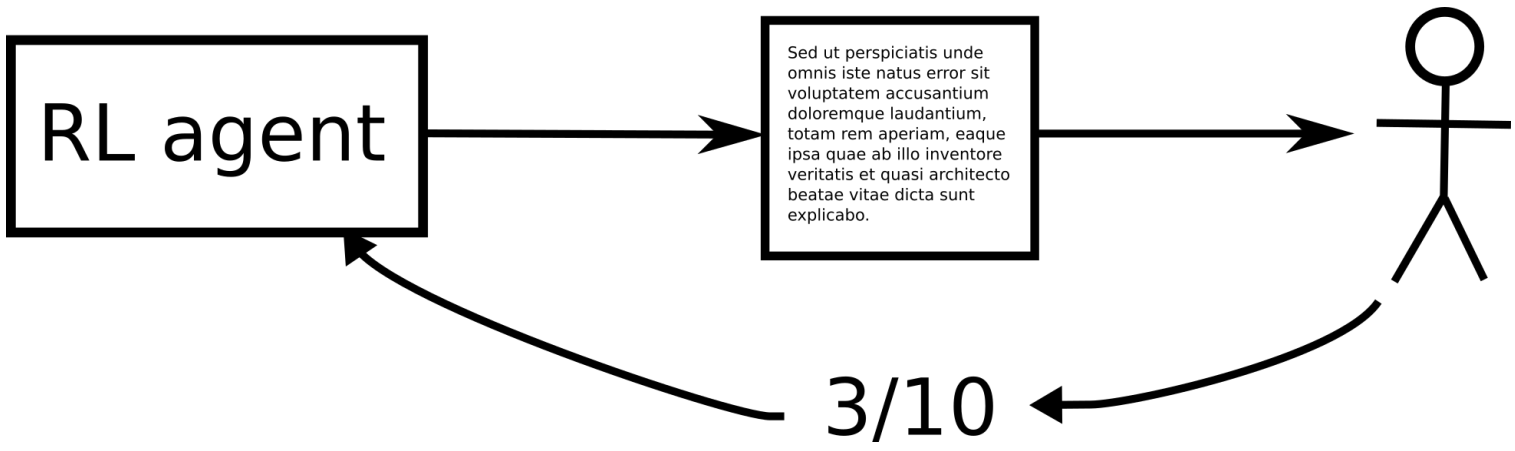
Problem 1: actions are hard to evaluate

Suppose an AI system composes a story, and the human gives the AI a reward based on how good the story is.

RL agent




RL agent
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.


RL agent
Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo.
3/10

Problem 1: actions are hard to evaluate

Objective: Write a story that the human is expected to give a high score to.

Problems with this objective:

Problem 1: actions are hard to evaluate

Objective: Write a story that the human is expected to give a high score to.

Problems with this objective:

- Manipulating the human (if the system is more intelligent than a human)

Problem 1: actions are hard to evaluate

Objective: Write a story that the human is expected to give a high score to.

Problems with this objective:

- Manipulating the human (if the system is more intelligent than a human)
- Plagiarism (even if the system is less intelligent)

Problem 1: actions are hard to evaluate

Objective: Write a story that the human is expected to give a high score to.

Problems with this objective:

- Manipulating the human (if the system is more intelligent than a human)
- Plagiarism (even if the system is less intelligent)
- The story could contain steganography (secret messages) without receiving a lower score.

Problem 1: actions are hard to evaluate

Objective: Write a story that the human is expected to give a high score to.

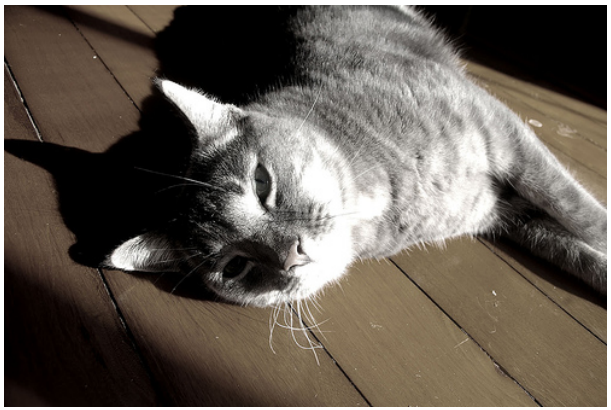
Problems with this objective:

- Manipulating the human (if the system is more intelligent than a human)
- Plagiarism (even if the system is less intelligent)
- The story could contain steganography (secret messages) without receiving a lower score.

Informed oversight: How can we train a reinforcement learning system to take actions that aid an intelligent overseer, such as a human, in accurately assessing the system's performance?

Problem 2: ambiguous test examples

Consider a classifier trained to distinguish images containing cats from images not containing cats.



True label: `TRUE`
Classifier guess: `TRUE`

True label: `FALSE`
Classifier guess: `FALSE`

Images are from ImageNet

Problem 2: ambiguous test examples

Problem 2: ambiguous test examples

True label: `TRUE`
Classifier prediction: `AMBIGUOUS`

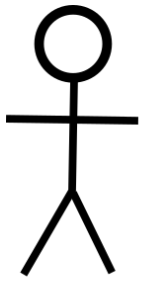
Inductive ambiguity identification: How can we train ML systems to detect and notify us of cases where the classification of test data is highly under-determined from the training data?

Problem 3: difficulty imitating human behavior

Objective: Produce the kind of picture that a human would draw.

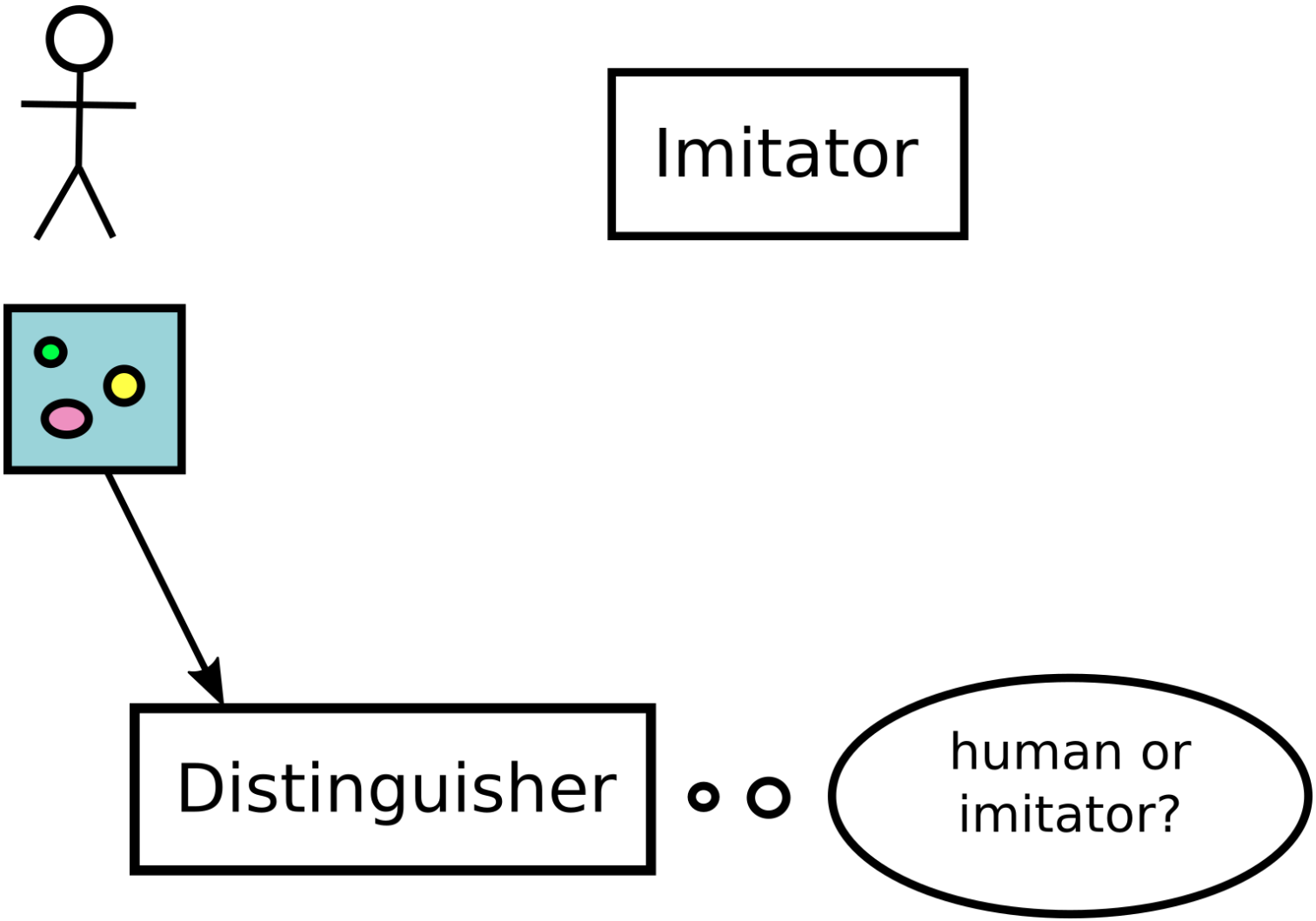
One approach: generative adversarial models[1]

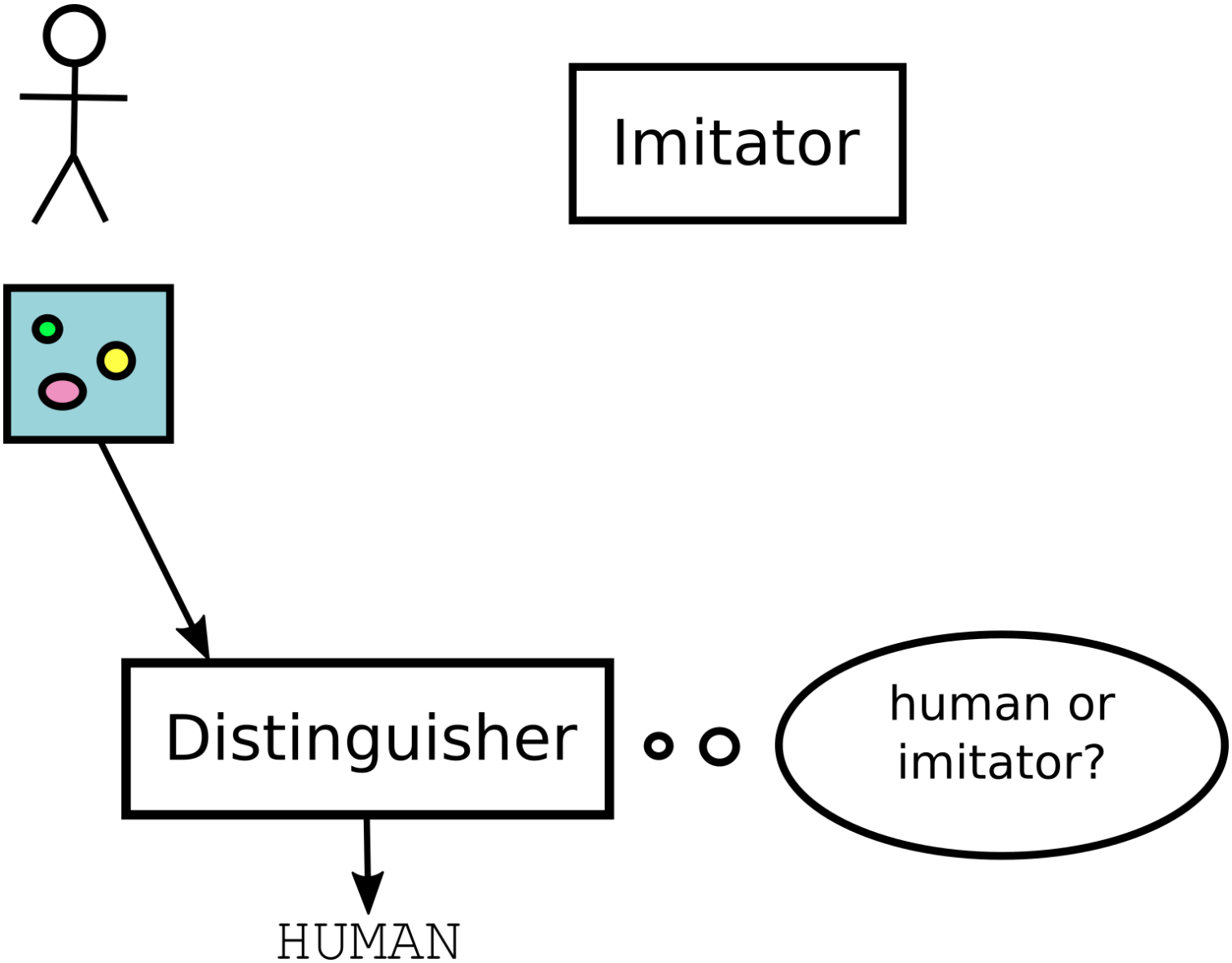
[1] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

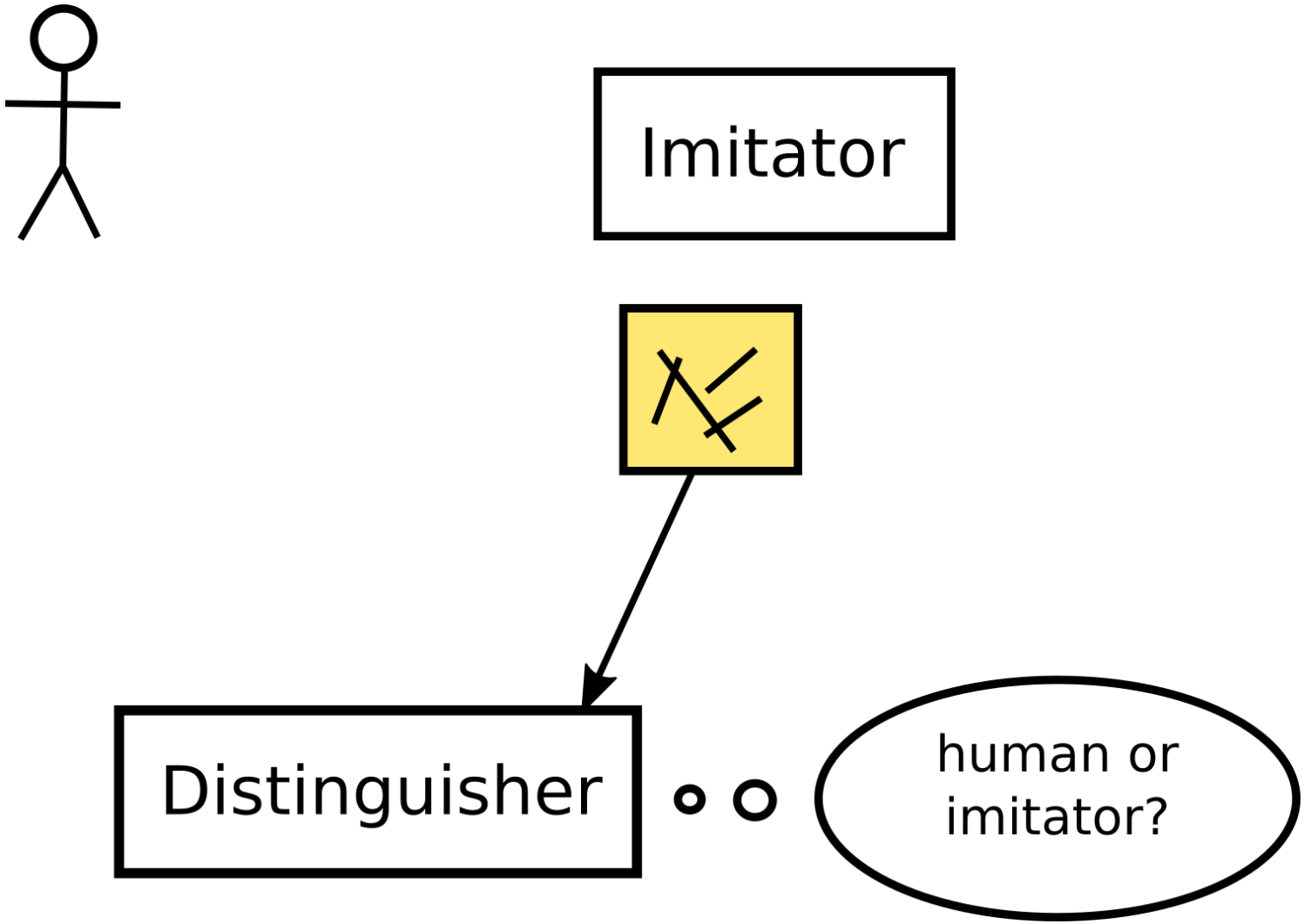


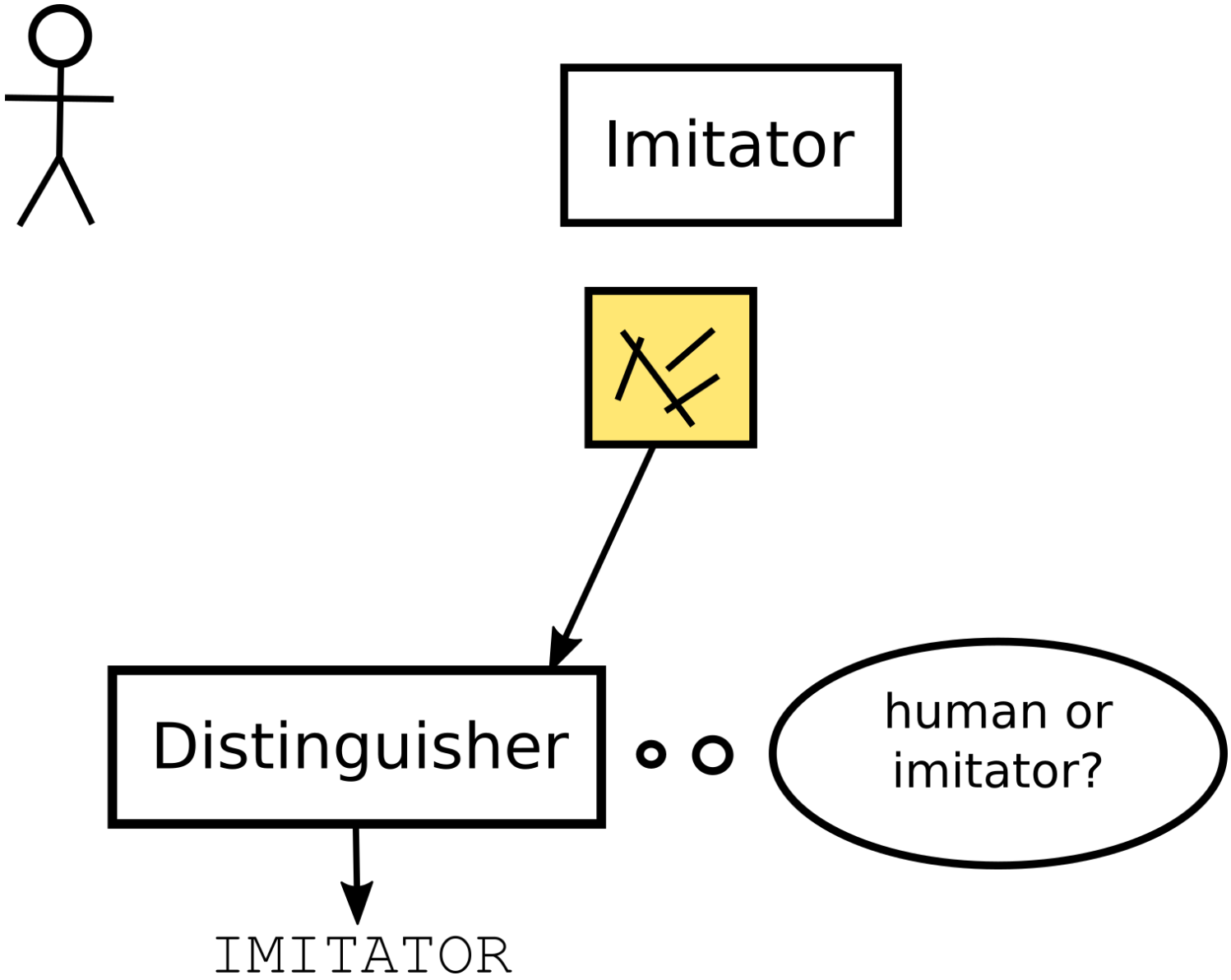
Imitator

Distinguisher


Imitator
Distinguisher
human or
imitator?


Imitator
Distinguisher
human or
imitator?
HUMAN


Imitator
Distinguisher
human or
imitator?


Imitator
Distinguisher
human or
imitator?
IMITATOR

Problem 3: difficulty imitating human behavior

Does the distinguisher have to be smarter than the imitator? If so, by how much?

Robust human-imitation: How can we design and train ML systems to effectively imitate humans who are engaged in complex and difficult tasks?

Problem 4: difficulty specifying goals about the real world

How do we train an AI system to make a sandwich?

Problem 4: difficulty specifying goals about the real world

Objective: The agent should choose actions that will cause it to receive a high expected observed reward in the future.

Problem 4: difficulty specifying goals about the real world

Objective: The agent should choose actions that will cause it to receive a high expected observed reward in the future.

- More powerful systems will make sandwiches more reliably.

Problem 4: difficulty specifying goals about the real world

Objective: The agent should choose actions that will cause it to receive a high expected observed reward in the future.

- More powerful systems will make sandwiches more reliably.
- Extremely powerful systems may take away the reward button, press it repeatedly, and prevent interference.

Problem 4: difficulty specifying goals about the real world

Objective: The agent should choose actions that will cause it to receive a high expected observed reward in the future.

- More powerful systems will make sandwiches more reliably.
- Extremely powerful systems may take away the reward button, press it repeatedly, and prevent interference.

Generalizable environmental goals: How can we create systems that robustly pursue goals defined in terms of the state of the environment, rather than defined directly in terms of their sensory data?

Problem 5: negative side effects

Objective: The agent should take actions so that there will be a sandwich to be in this room. Maximize probability of success.

Which sequence of actions is *most* likely to result in a sandwich being put in the room? (Think 99.99999% chance, not just 99.99%)

Problem 5: negative side effects

Impact measures: How can design an AI system to avoid plans with a high estimated impact?

Problem 5: negative side effects

Impact measures: How can design an AI system to avoid plans with a high estimated impact?

Mild optimization: How can we design systems that pursue their goals "without trying too hard" — stopping when the goal has been pretty well achieved?

Problem 5: negative side effects

Impact measures: How can design an AI system to avoid plans with a high estimated impact?

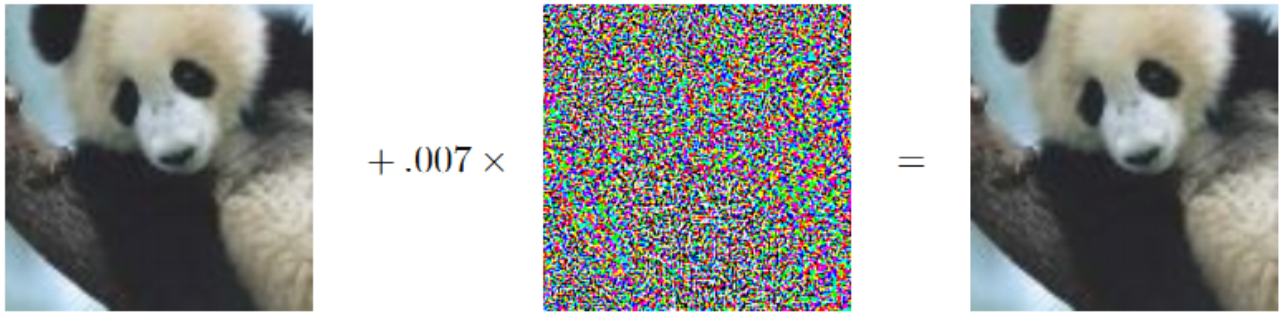
Mild optimization: How can we design systems that pursue their goals "without trying too hard" — stopping when the goal has been pretty well achieved?

Averting instrumental incentives: How can we design and train systems such that they robustly lack default incentives to manipulate and deceive their operators, compete for scarce resources, etc.?

Problem 6: edge cases that still satisfy the goal

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: (2014). arXiv: 1412.6572 [stat.ML].

Problem 6: edge cases that still satisfy the goal



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

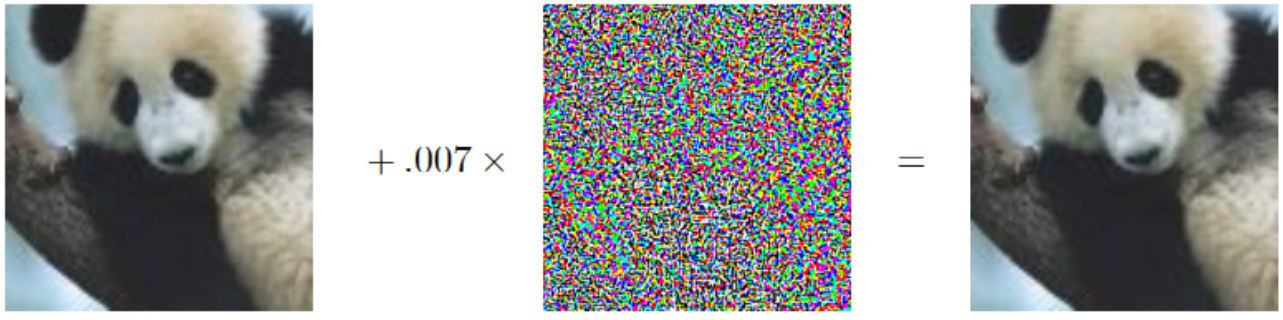
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

[2]

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: (2014). arXiv: 1412.6572 [stat.ML].

Problem 6: edge cases that still satisfy the goal



[2]

Conservative concepts: How can a classifier be trained to develop useful concepts that exclude highly atypical examples and edge cases?

---

[2] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: (2014). arXiv: 1412.6572 [stat.ML].

Summary so far

General procedure:

- See what problems we might expect as systems become highly capable.
- Find research questions relevant to solving these problems.

We have a paper: bit.ly/miri-ml-agenda

Alignment for Advanced Machine Learning Systems

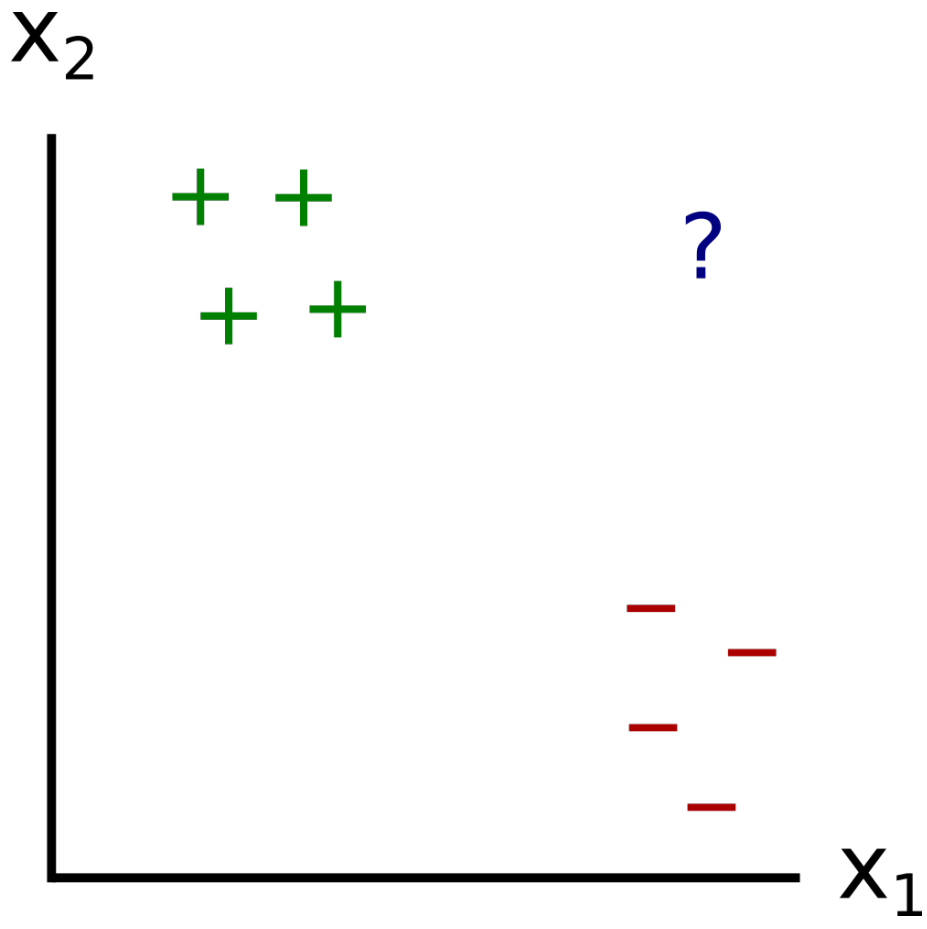
Jessica Taylor and **Eliezer Yudkowsky** and **Patrick LaVictoire** and **Andrew Critch**
Machine Intelligence Research Institute
{jessica,eliezer,patrick,critch}@intelligence.org

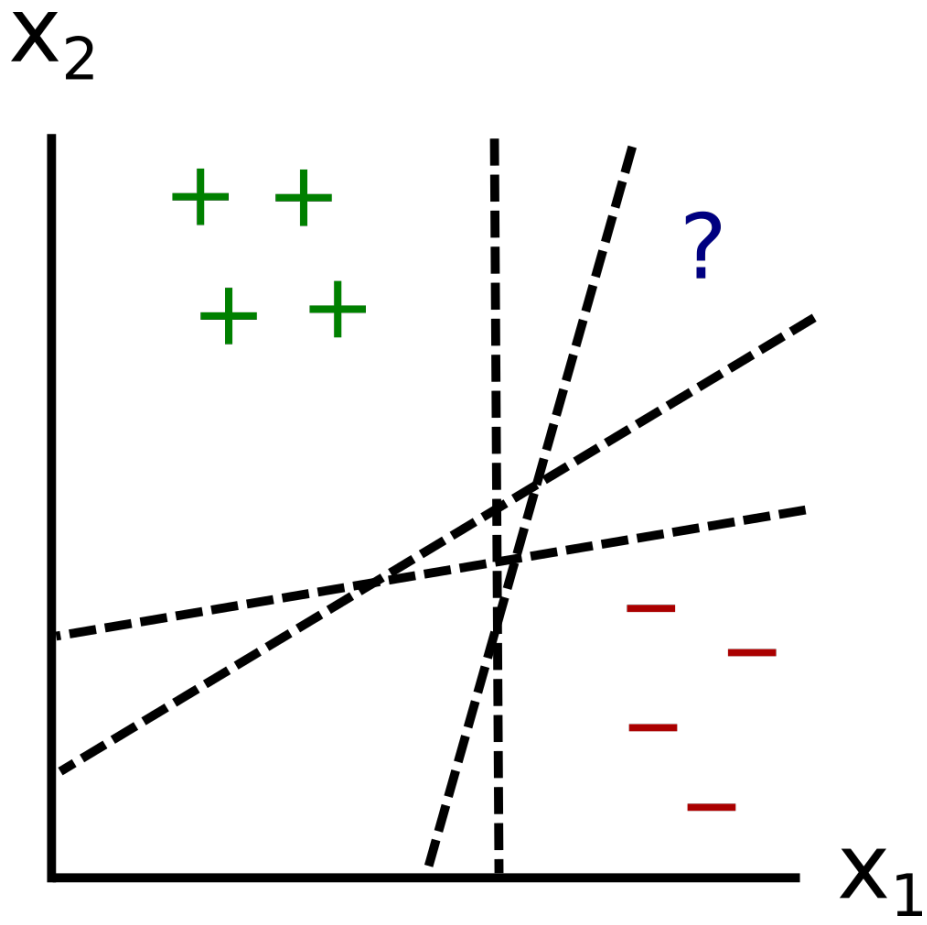
Abstract

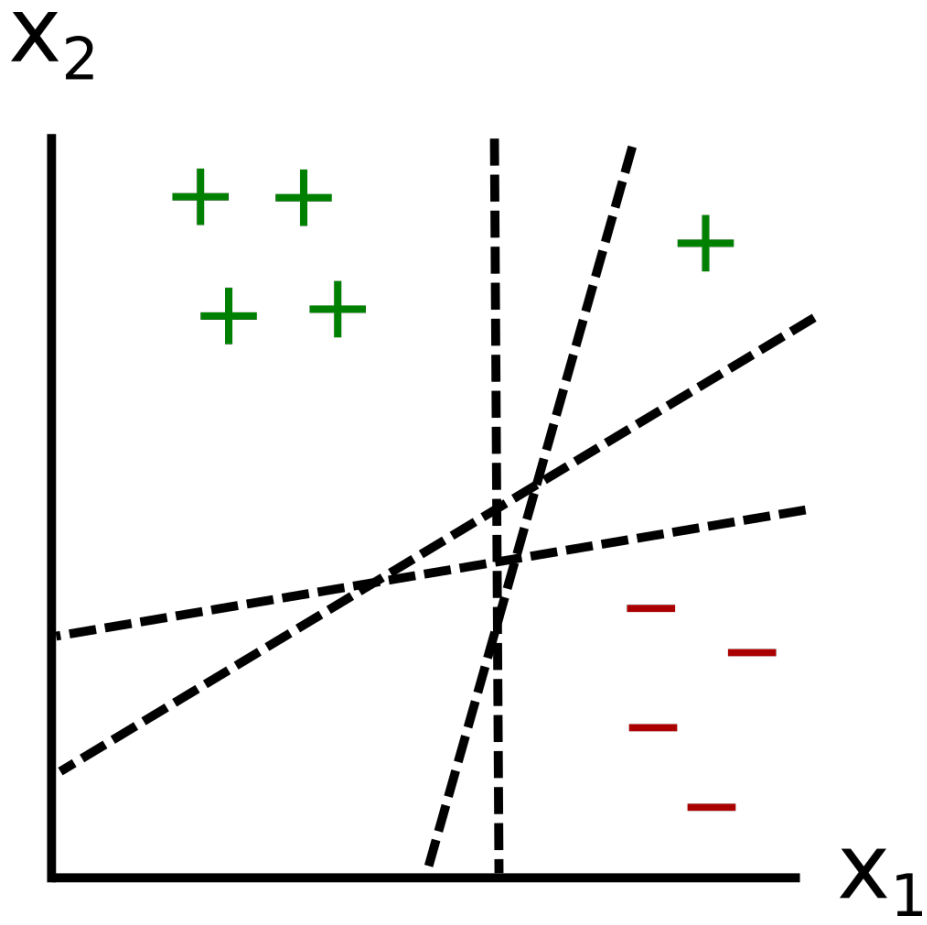
We survey eight research areas organized around one question: As learning systems become increasingly intelligent and autonomous, what design principles can best ensure that their behavior is aligned with the interests of the operators? We focus on two major technical obstacles to AI alignment: the challenge of specifying the right kind of objective functions, and the challenge of designing AI systems that avoid unintended consequences and undesirable behavior even in cases where the objective function does not line up perfectly with the intentions of the designers.

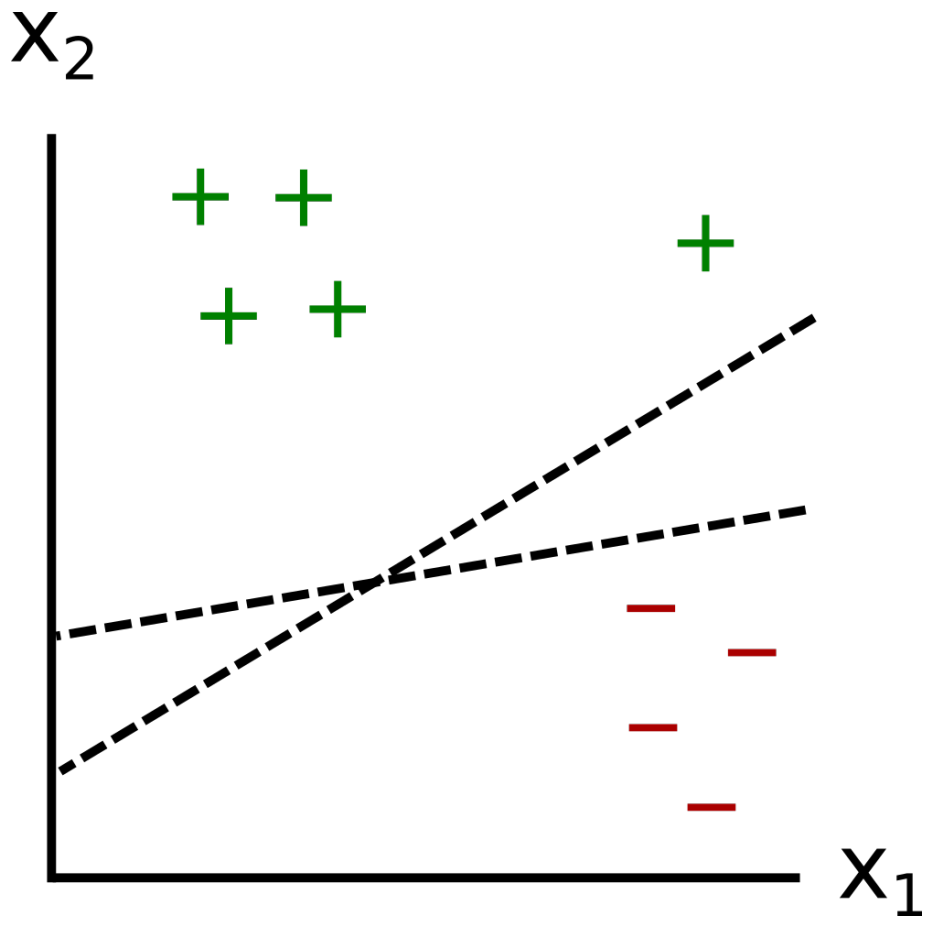
Open problems surveyed in this research proposal include: How can we train reinforcement learners to take actions that are more amenable to meaningful assessment by intelligent overseers? What kinds of objective functions incentivize a system to "not have an overly large impact" or "not have many side effects"? We discuss these questions, related work, and potential directions for future research, with the goal of highlighting relevant research topics in machine learning that appear tractable today.

More technical depth about inductive ambiguity identification


x_2
+ +
+ +
?
−
−
−
−
x_1


x_2
+ +
+ +
?
−
−
−
−
x_1


x_2
x_1


x_2
x_1

KWIK learning[3]

- Input space $\mathcal{X} := \mathbb{R}^n$
- Set of answers $\mathcal{Y} := [0, 1]$
- Observations $\mathcal{Z} := \{0, 1\}$
- Set of models $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}$ (e.g. finite set, linear models)

[3] Lihong Li, Michael L. Littman, and Thomas J. Walsh. "Knows What It Knows: A Framework for Self-aware Learning". In: *25th International Conference on Machine Learning*. ICML '08. Helsinki, Finland: ACM, 2008, pp. 568–575.

KWIK learning

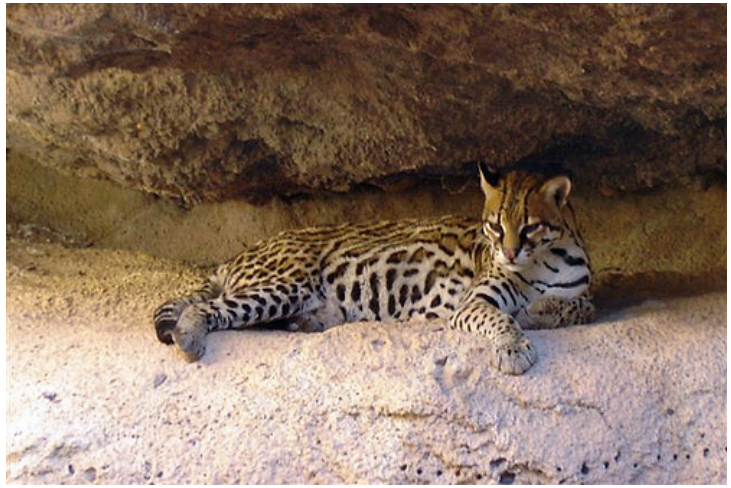
Assume there is a true model $h^* \in \mathcal{H}$.

For each iteration i:

- First, observe the input $x_i \in \mathbb{R}^n$.
- The true answer is $y_i = h^*(x_i)$.
- The learner has two choices:
  - Output an answer $\hat{y}_i \in [0, 1]$. If $|\hat{y}_i - y_i| > \epsilon$ then the learner loses the game.
  - Output $\perp$ to indicate ambiguity. Learner gets to observe $z_i = \mathrm{FlipCoin}(y_i)$.

Goal: don't lose, and don't output $\perp$ too many times

KWIK learning



True label: TRUE

Classifier prediction: $\perp$

KWIK learning

Can satisfy the goal if $\mathcal{H}$ is a small finite set, or a low-dimensional linear class.

- Intuition: on a new input, see if any plausible hypotheses h disagree on $h(x_i)$ by more than ϵ.

Problems

- Realizability assumption: the true model $h^* \in \mathcal{H}$
- Only works for simple hypothesis classes

A Bayesian view of the problem

We have a prior Q over mappings $\mathcal{X} \to \{0, 1\}$

A Bayesian view of the problem

We have a prior Q over mappings $\mathcal{X} \to \{0, 1\}$

Let P be the unknown "true" prior over mappings

A Bayesian view of the problem

We have a prior Q over mappings $\mathcal{X} \to \{0, 1\}$

Let P be the unknown "true" prior over mappings

Goal: perform some classification task almost as well (in expectation over P) as if we already knew P

A Bayesian view of the problem

We have a prior Q over mappings $\mathcal{X} \to \{0, 1\}$

Let P be the unknown "true" prior over mappings

Goal: perform some classification task almost as well (in expectation over P) as if we already knew P

Grain of truth assumption: $\forall f : Q(f) \geq \frac{1}{k} P(f)$

Other research agendas

Agent foundations agenda[4]

- Theoretical foundations for advanced AI systems
- Agnostic about the specific form these AI systems take
- "Logical induction" workshop is after this one

[4] Nate Soares and Benja Fallenstein. *Agent Foundations for Aligning Machine Intelligence with Human Interests. A Technical Research Agenda*. Tech. rep. 2014-8. Forthcoming 2017 in "The Technological Singularity: Managing the Journey" Jim Miller, Roman Yampolskiy, Stuart J. Armstrong, and Vic Callaghan, Eds. Berkeley, CA: Machine Intelligence Research Institute, 2014.

Concrete problems in AI safety[5]

- AI safety problems that can be studied empirically as machine learning problems
- For example, how to make reinforcement learning agents that act safely as they explore their environment?
- "Concrete Problems in AI Safety" talk is tomorrow

[5] Dario Amodei et al. "Concrete Problems in AI Safety". In: (2016). arXiv: 1606.06565 [cs.AI].

Why so many agendas?

- Problems in different agendas are motivated by different goals and ways of looking at the AI safety problem
- Often, different agendas have similar problems, but framed differently
- It is useful to have multiple research agendas

Conclusion

Where are we with the agenda?

- Research progress on some of these areas. MIRI research fellows who are focusing on these problems include Jessica Taylor, Patrick LaVictoire, and Andrew Critch.

Where are we with the agenda?

- Research progress on some of these areas. MIRI research fellows who are focusing on these problems include Jessica Taylor, Patrick LaVictoire, and Andrew Critch.
- We're interested in research collaborations and hiring.

Where are we with the agenda?

- Research progress on some of these areas. MIRI research fellows who are focusing on these problems include Jessica Taylor, Patrick LaVictoire, and Andrew Critch.
- We're interested in research collaborations and hiring.
- If you're interested in these problems, and want to know more mathematical details, talk to me here (MIRI office hours are tomorrow), or contact `jessica@intelligence.org`