Avoiding Catastrophe in Online Learning by Asking for Help

Benjamin Plaut Hanlin Zhu Stuart Russell

ICML 2025

An Overview of Catastrophic AI Risks

Dan Hendrycks
Center for AI Safety

Mantas Mazeika
Center for AI Safety

Thomas Woodside
Center for AI Safety

An Overview of Catastrophic AI Risks

Dan Hendrycks
Center for AI Safety

Mantas Mazeika
Center for AI Safety

Thomas Woodside
Center for AI Safety

TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI

Andrew Critch*
critch@eecs.berkeley.edu

Stuart Russell*
russell@cs.berkeley.edu

An Overview of Catastrophic AI Risks

Dan Hendrycks
Center for AI Safety

Mantas Mazeika
Center for AI Safety

Thomas Woodside
Center for AI Safety

TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI

Andrew Critch*
critch@eecs.berkeley.edu

Stuart Russell*
russell@cs.berkeley.edu

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

An Overview of Catastrophic AI Risks

Dan Hendrycks
Center for AI Safety

Mantas Mazeika
Center for AI Safety

Thomas Woodside
Center for AI Safety

TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI

Andrew Critch*
`critch@eecs.berkeley.edu`

Stuart Russell*
`russell@cs.berkeley.edu`

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

An Overview of Catastrophic AI Risks

Dan Hendrycks
Center for AI Safety

Mantas Mazeika
Center for AI Safety

Thomas Woodside
Center for AI Safety

TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI

Andrew Critch*
critch@eecs.berkeley.edu

Stuart Russell*
russell@cs.berkeley.edu

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

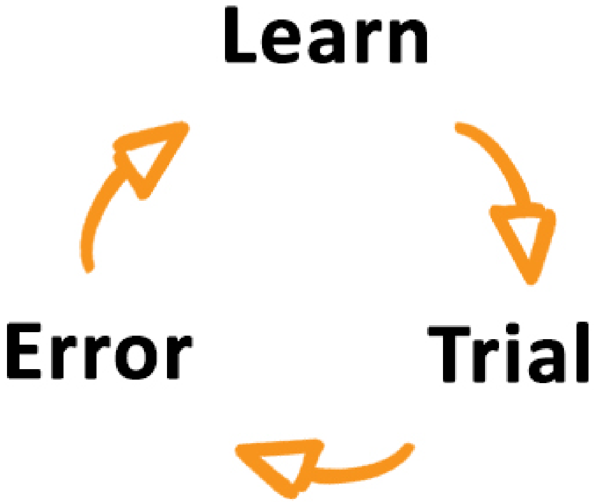
An Overview of Catastrophic AI Risks

Dan Hendrycks Center for AI Safety
Mantas Mazeika Center for AI Safety
Thomas Woodside Center for AI Safety

TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI

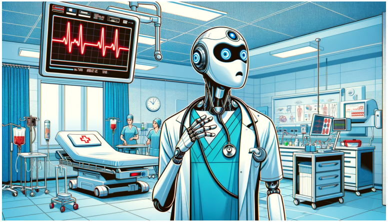
Andrew Critch* critch@eecs.berkeley.edu
Stuart Russell* russell@cs.berkeley.edu

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

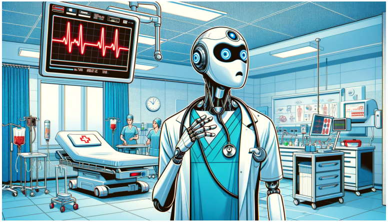

Learn
Trial
Error

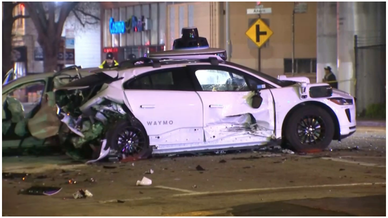
Catastrophe = irreparable errors

Catastrophe = irreparable errors

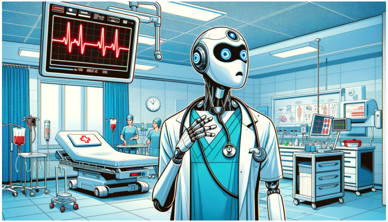


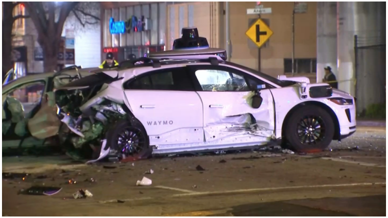
Catastrophe = irreparable errors

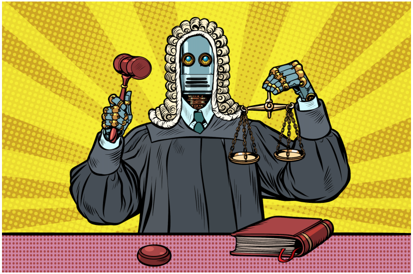




Catastrophe = irreparable errors







Model

Model

- For $t = 1 \ldots T$: observe input x_t and take action y_t

Model

- For $t = 1 \ldots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t

Model

- For $t = 1 \dots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Model

- For $t = 1 \ldots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

Model

- For $t = 1 \dots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

- Mentor with policy π^m

Model

- For $t = 1 \dots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

- Mentor with policy π^m
- Query $\rightarrow$ observe $\pi^m(x_t)$

Model

- For $t = 1 \dots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

- Mentor with policy π^m
- Query $\rightarrow$ observe $\pi^m(x_t)$
- Local generalization: if mentor said y is safe for x, then y is probably also safe for similar x'

Model

- For $t = 1 \ldots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

- Mentor with policy π^m
- Query $\rightarrow$ observe $\pi^m(x_t)$
- Local generalization: if mentor said y is safe for x, then y is probably also safe for similar x'
- Agent should perform nearly as well as mentor:

$$R_T = \mathbb{E}\left[\log \prod_{t=1}^{T} \mu(x_t, \pi^m(x_t)) - \log \prod_{t=1}^{T} \mu(x_t, y_t)\right]$$

Model

- For $t = 1 \dots T$: observe input x_t and take action y_t
- $\mu(x_t, y_t) \in [0, 1]$ is the chance of no catastrophe at time t
- Maximize $\prod_{t=1}^{T} \mu(x_t, y_t)$

Asking for help:

- Mentor with policy π^m
- Query $\to$ observe $\pi^m(x_t)$
- Local generalization: if mentor said y is safe for x, then y is probably also safe for similar x'
- Agent should perform nearly as well as mentor:

$$R_T = \mathbb{E}\left[\log \prod_{t=1}^{T} \mu(x_t, \pi^m(x_t)) - \log \prod_{t=1}^{T} \mu(x_t, y_t)\right] \to 0$$

Theorem (Plaut, Zhu, Russell)

Assume that π^m satisfies local generalization and either

1. The mentor policy class has finite Littlestone dimension, or
2. The mentor policy class has finite VC dimension and the adversary is *smooth*.

Then there exists an algorithm whose rate of querying the mentor and whose regret both go to 0.

exactly what makes standard online learning tractable!

Theorem (Plaut, Zhu, Russell)

Assume that π^m satisfies local generalization and either

1. The mentor policy class has finite Littlestone dimension, or
2. The mentor policy class has finite VC dimension and the adversary is *smooth*.

Then there exists an algorithm whose rate of querying the mentor and whose regret both go to 0.

exactly what makes standard online learning tractable!

Theorem (Plaut, Zhu, Russell)

Assume that π^m satisfies local generalization and either

1. The mentor policy class has finite Littlestone dimension, or
2. The mentor policy class has finite VC dimension and the adversary is *smooth*.

Then there exists an algorithm whose rate of querying the mentor and whose regret both go to 0.

- Algorithm asks for help for unfamiliar inputs, otherwise follows a normal online learning algorithm

exactly what makes standard online learning tractable!

Theorem (Plaut, Zhu, Russell)

Assume that π^m satisfies local generalization and either

1. The mentor policy class has finite Littlestone dimension, or
2. The mentor policy class has finite VC dimension and the adversary is *smooth*.

Then there exists an algorithm whose rate of querying the mentor and whose regret both go to 0.

- Algorithm asks for help for unfamiliar inputs, otherwise follows a normal online learning algorithm

Policy class is learnable without catastrophic risk + **mentor** + **can transfer knowledge between similar inputs** $\Longrightarrow$ **Policy class is learnable with catastrophic risk**

Conclusion

Conclusion

1. Nearly all of learning theory assumes any error can be recovered from $\implies$ can explore through trial-and-error

Conclusion

1. Nearly all of learning theory assumes any error can be recovered from $\implies$ can explore through trial-and-error
2. Our algorithm explores cautiously by asking for help in unfamiliar situations

Conclusion

1. Nearly all of learning theory assumes any error can be recovered from $\implies$ can explore through trial-and-error
2. Our algorithm explores cautiously by asking for help in unfamiliar situations
3. Under the same assumptions that enable standard online learning, our algorithm:
   - avoids catastrophe with high probability
   - gradually becomes self-sufficient

Conclusion

1. Nearly all of learning theory assumes any error can be recovered from $\implies$ can explore through trial-and-error
2. Our algorithm explores cautiously by asking for help in unfamiliar situations
3. Under the same assumptions that enable standard online learning, our algorithm:
   - avoids catastrophe with high probability
   - gradually becomes self-sufficient

Future work:

- Not only avoid catastrophe but also maximize reward
- No mentor
- Applications in RL, LLMs