Google Research

Chasing my dreams to communicate with machines.

Been Kim

CS224N @ Stanford March 2023

All opinions are my own.

Includes work with a lot of amazing folks I got to work with inside and outside of Google Brain team.

LLMs, Generative models are exciting,
And maybe… a little bit frightening.

Ultimately, we want this technologies to benefit “us”.

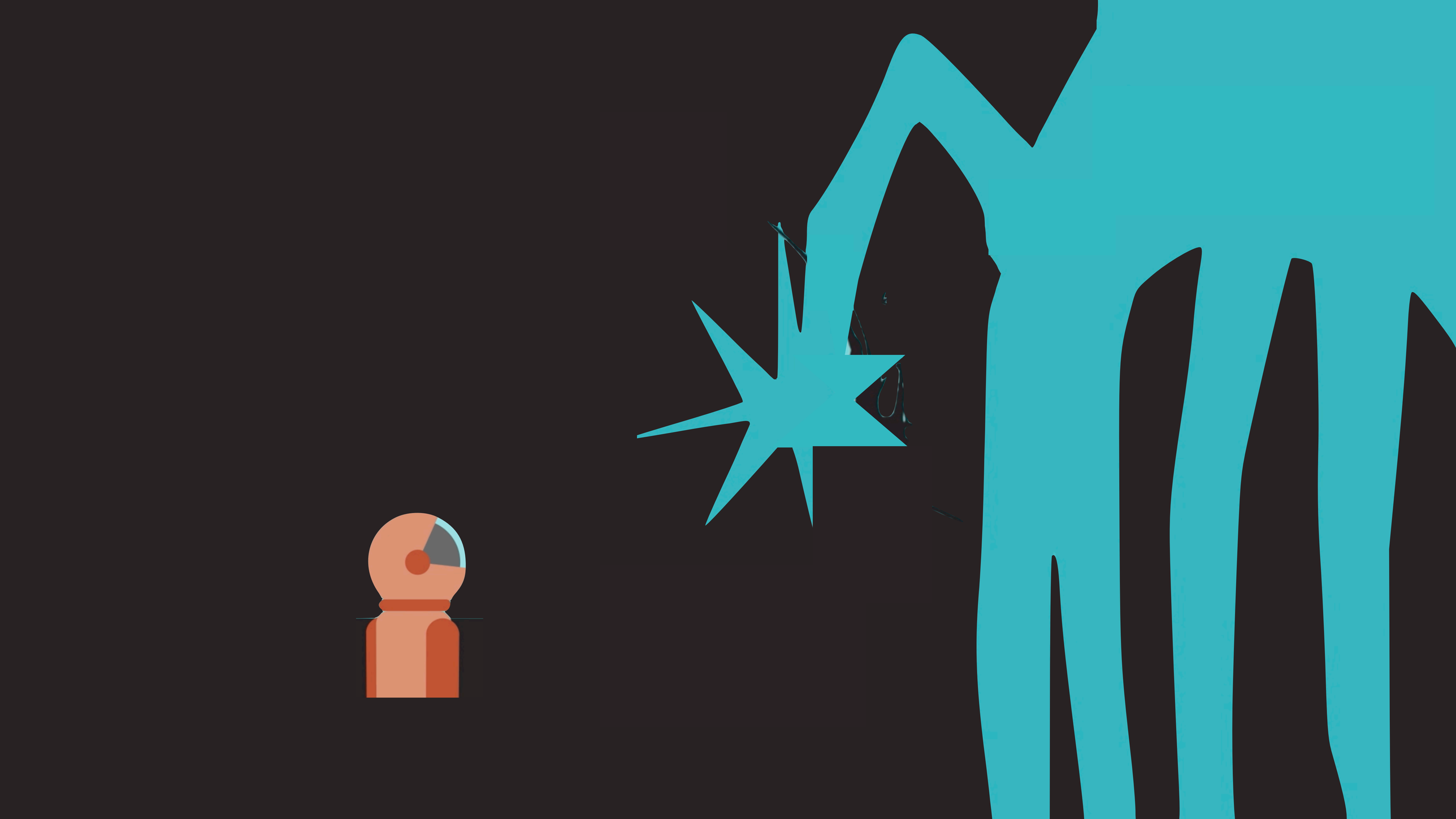
(my son will keep me accountable)

LLMs, Generative models are exciting,
And maybe… a little bit frightening.

Ultimately, we want this technologies to benefit “us”.

(my son will keep me accountable)

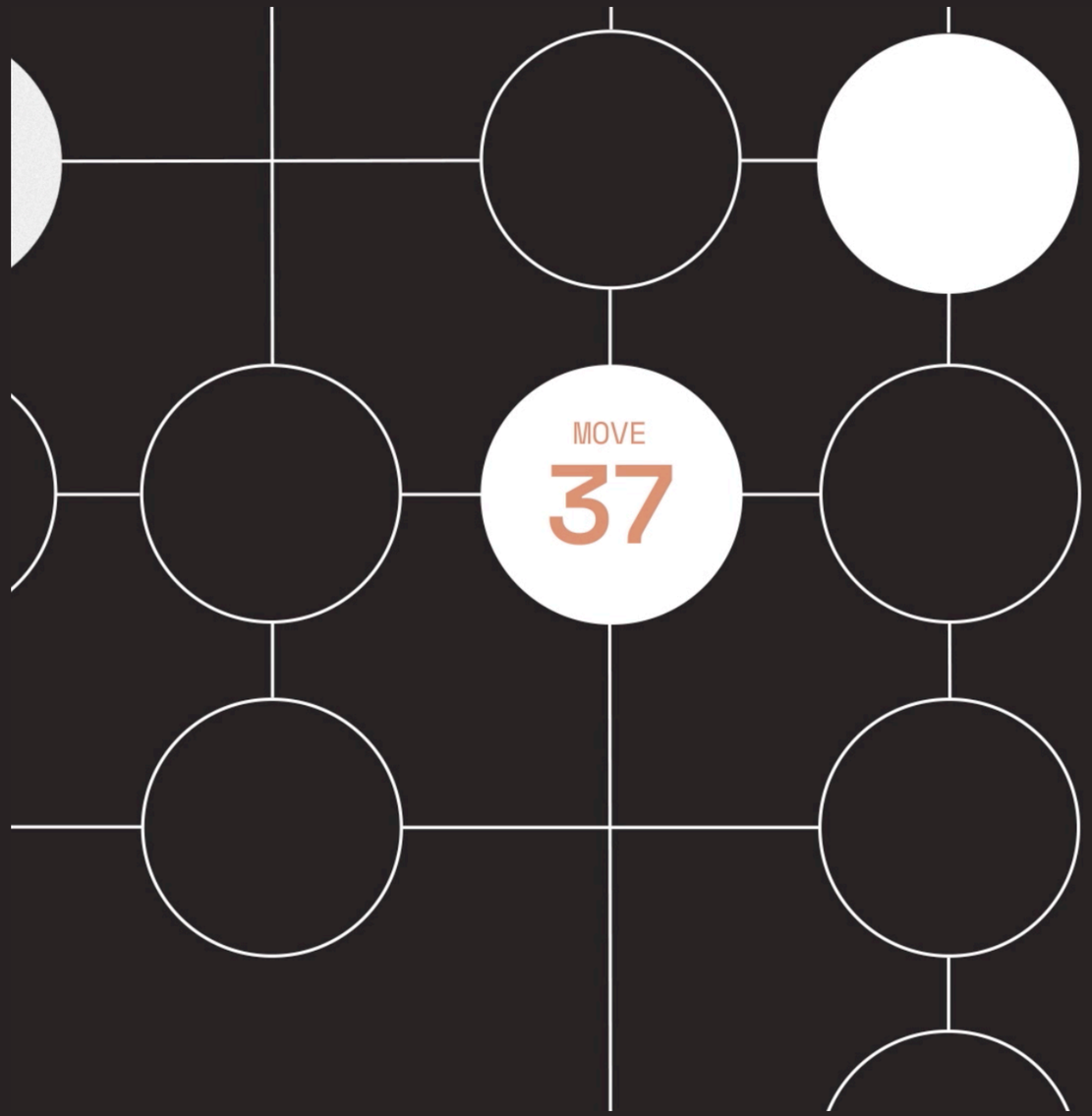
Why not optimize for what we want directly from the start?



My dreams:

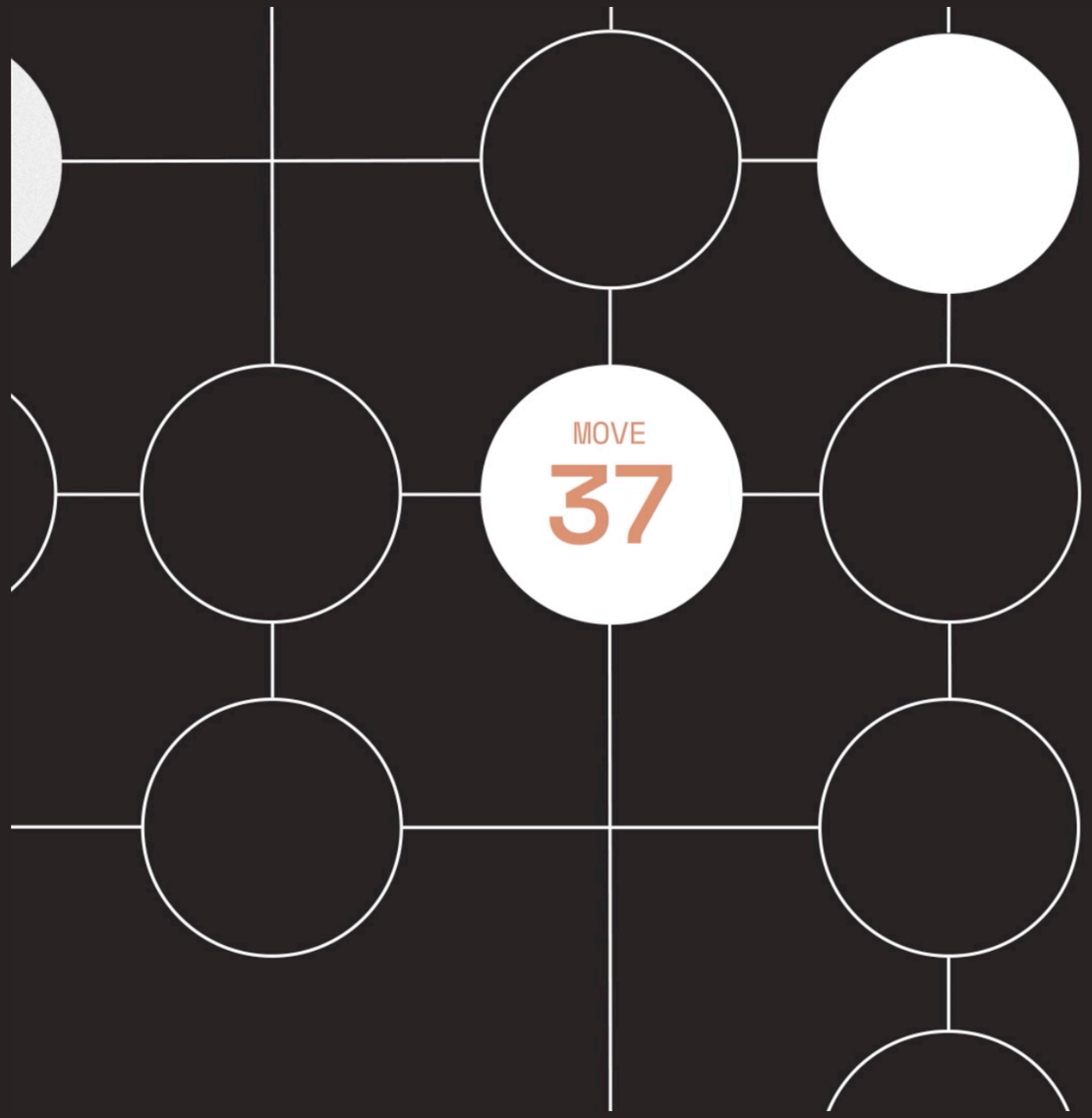
VS

ALPHA GO



“...that’s a very strange move.”

-Nine Dan Go Player, Commentator



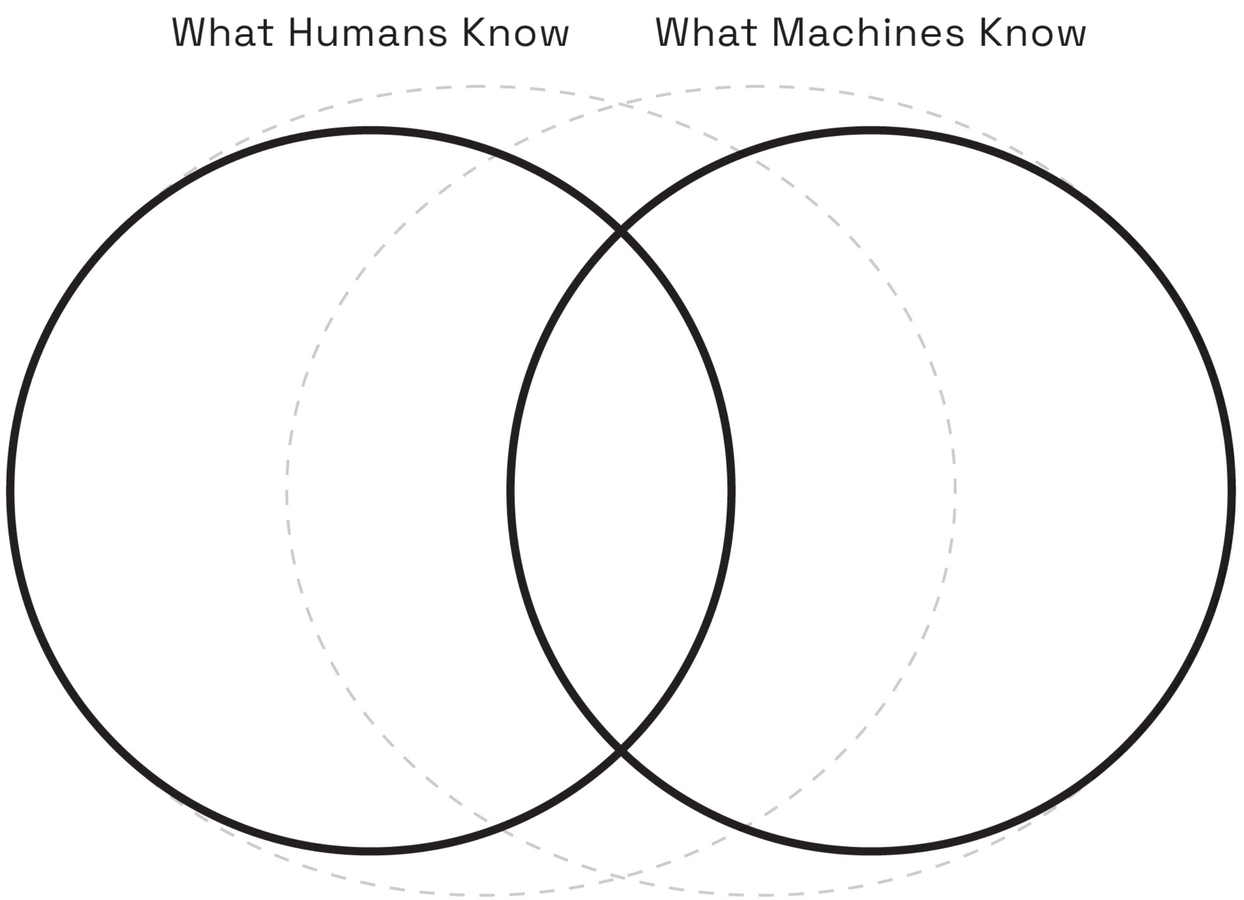
“...that’s a very strange move.”

-Nine Dan Go Player, Commentator

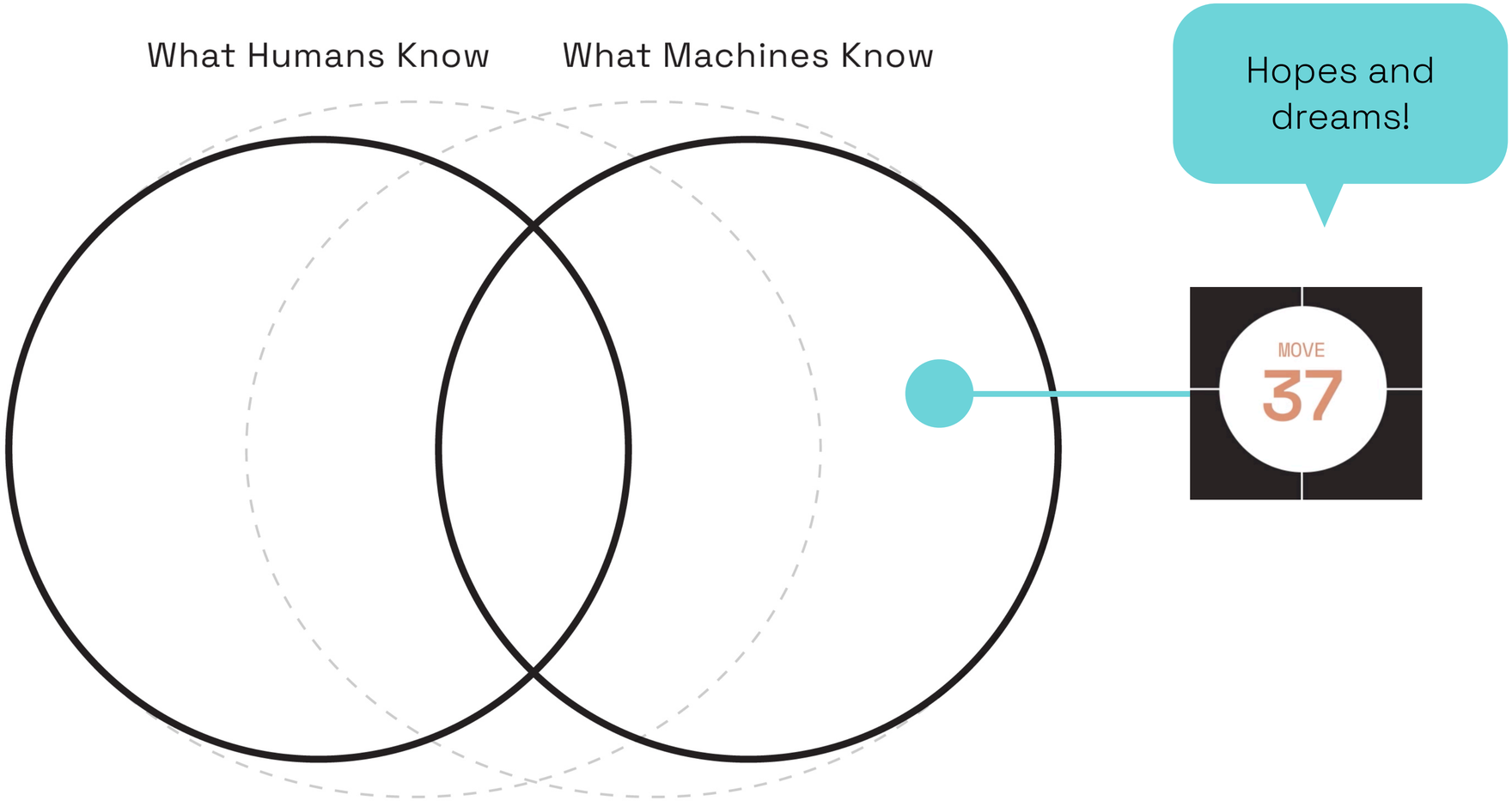
My hopes and dreams:

Learning something new by communicating with machines.

Human and Machine’s representational spaces

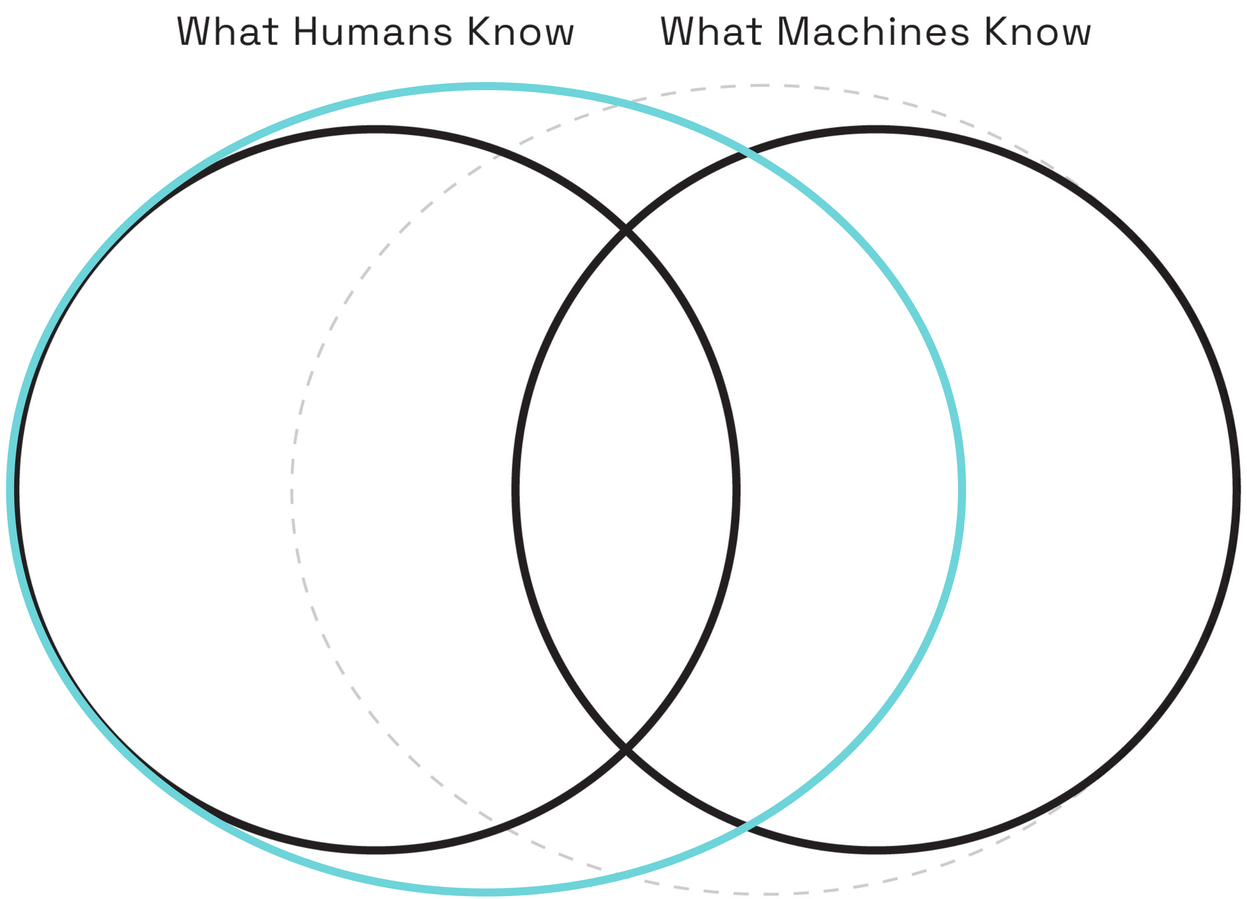


Human and Machine's representational spaces



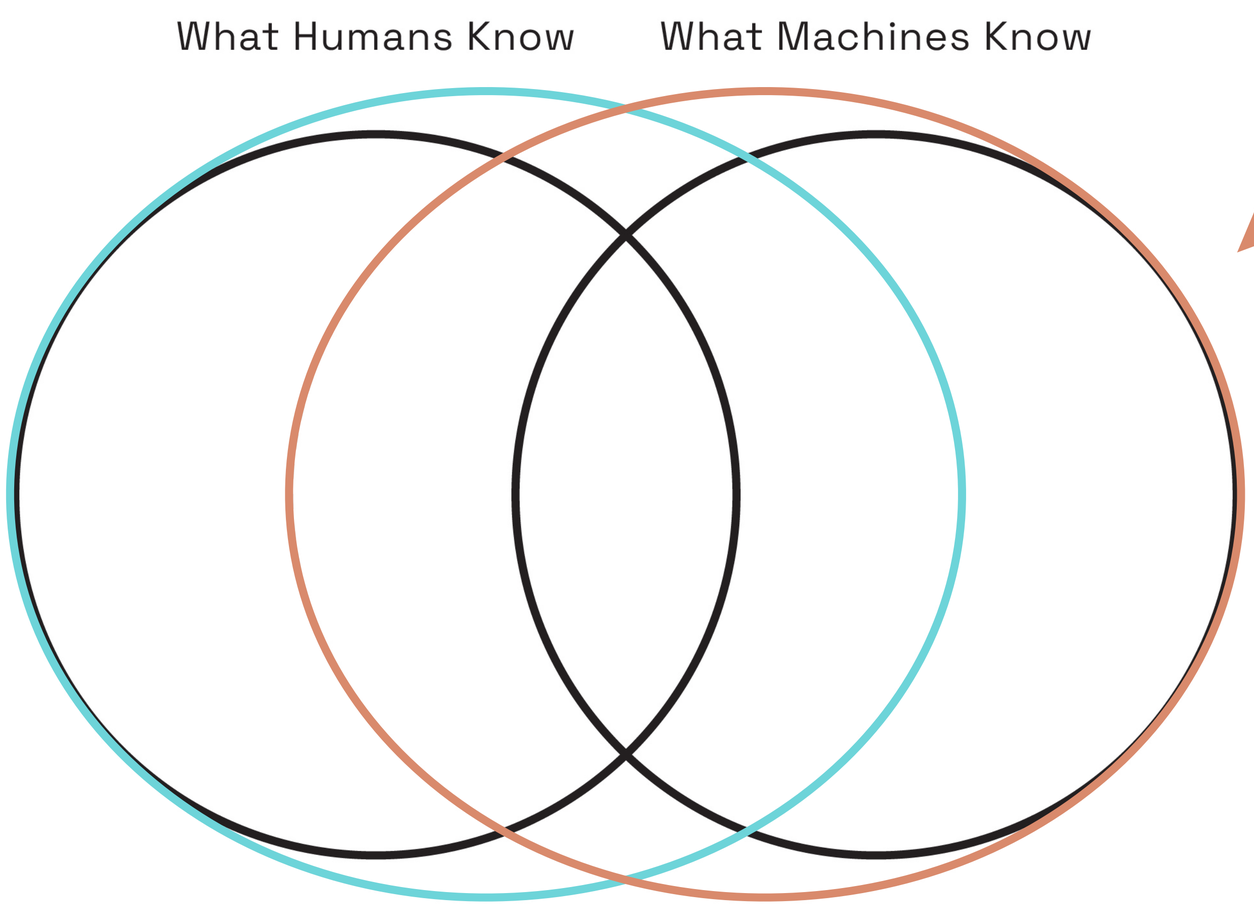
Human and Machine's representational spaces

Expanding what we know



Human and Machine's representational spaces

Expanding what we know



Build machines that better communicates with us

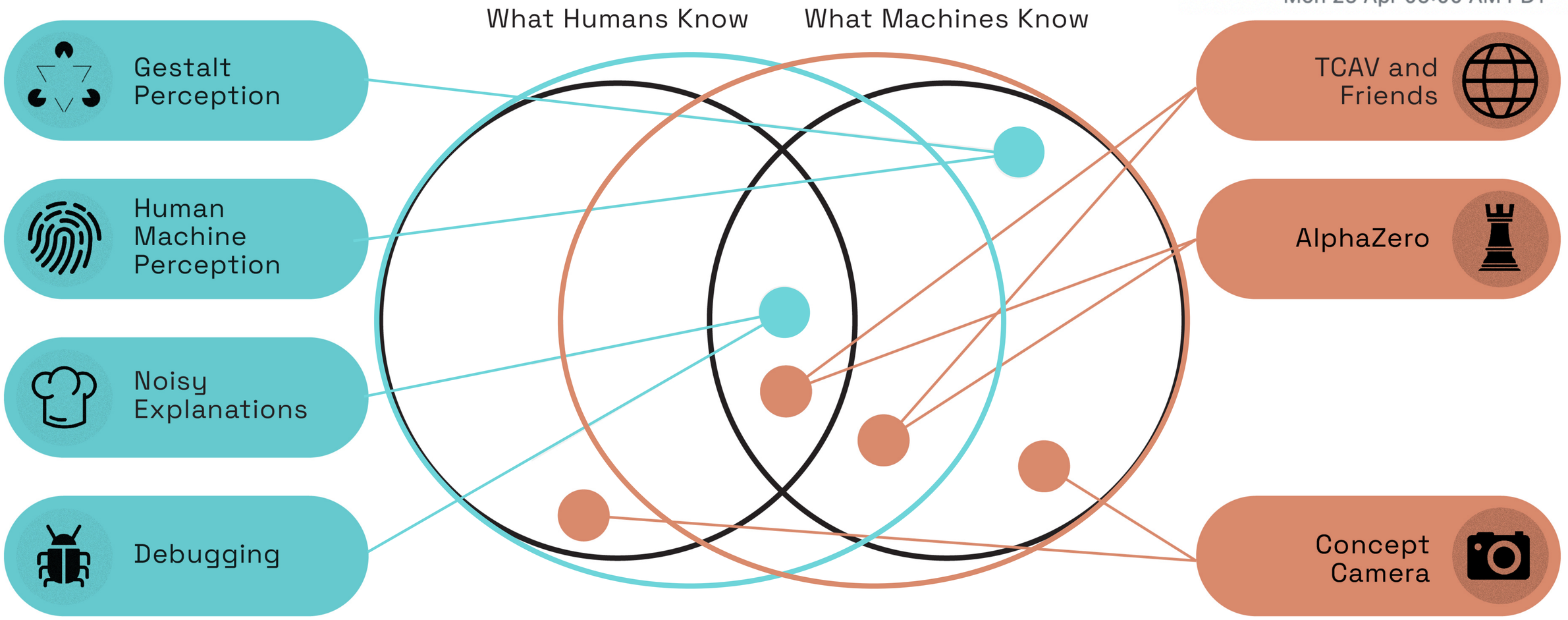
More on this: ICLR2022 Keynote

ICLR
Keynote

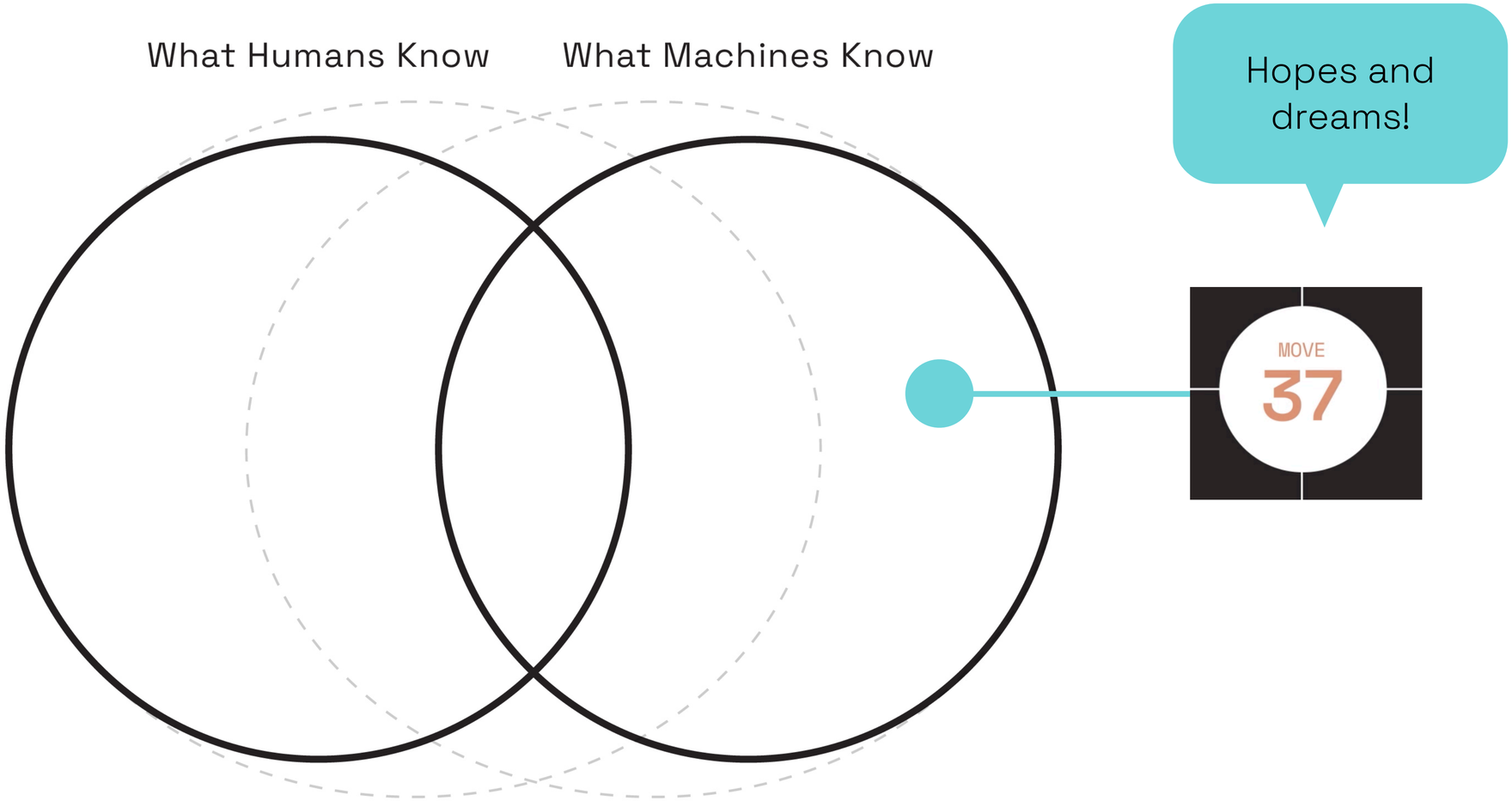
Beyond interpretability: developing a language to shape our relationships with AI

Been Kim

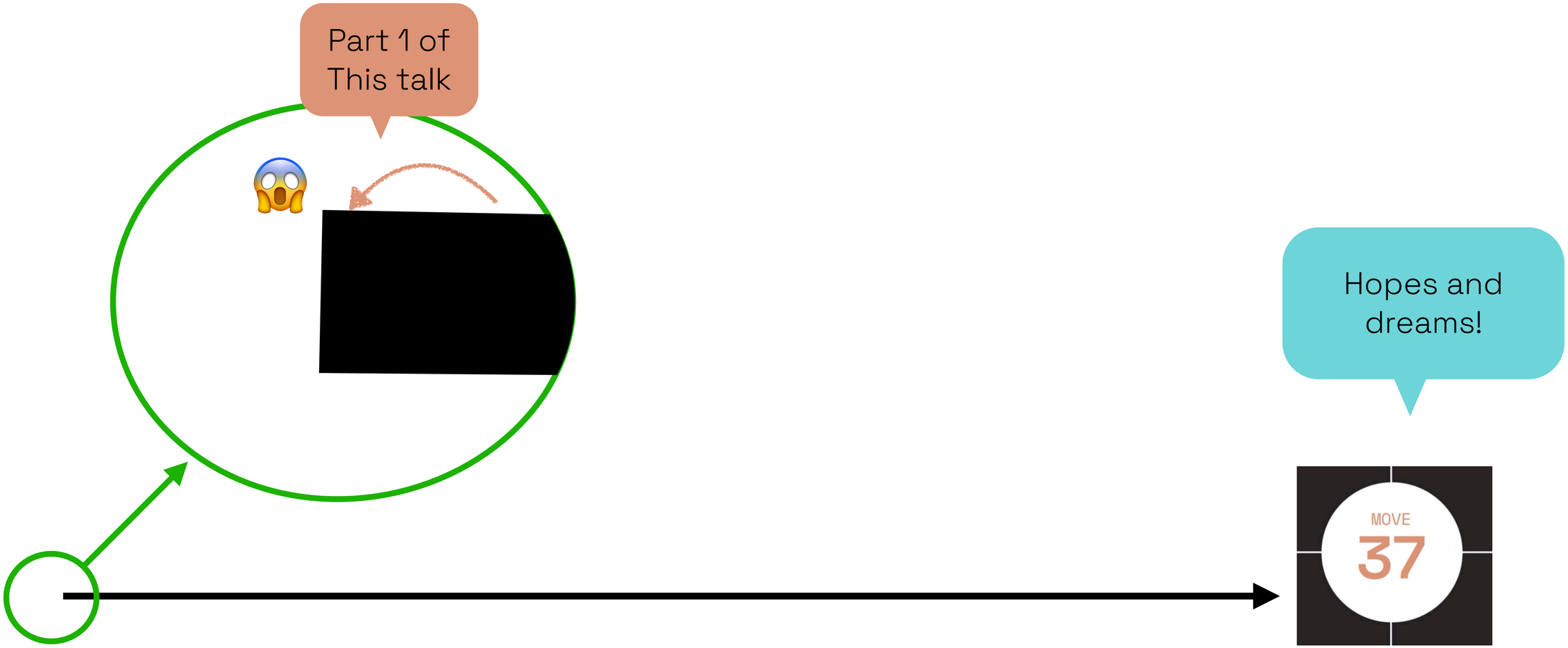
Mon 25 Apr 09:00 AM PDT

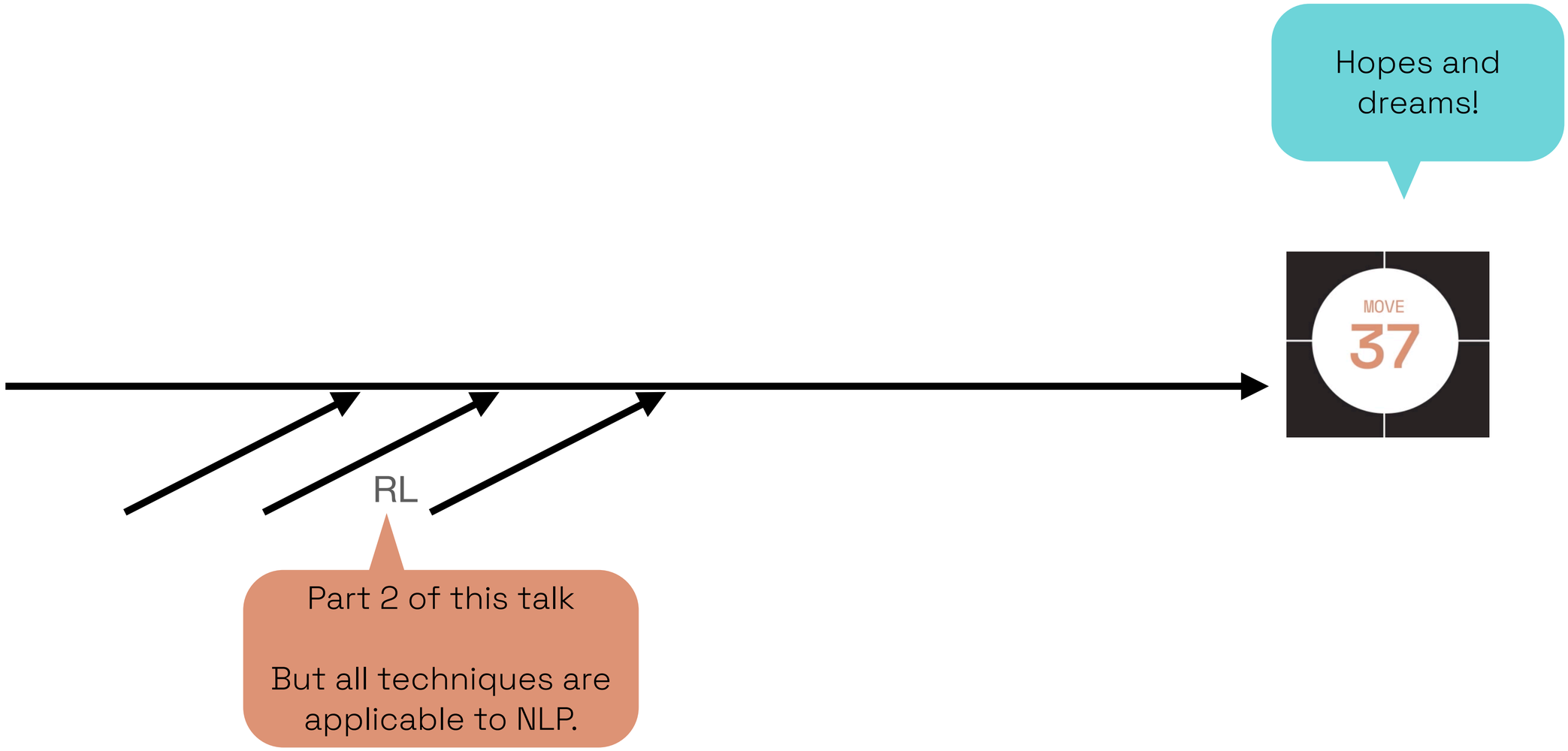


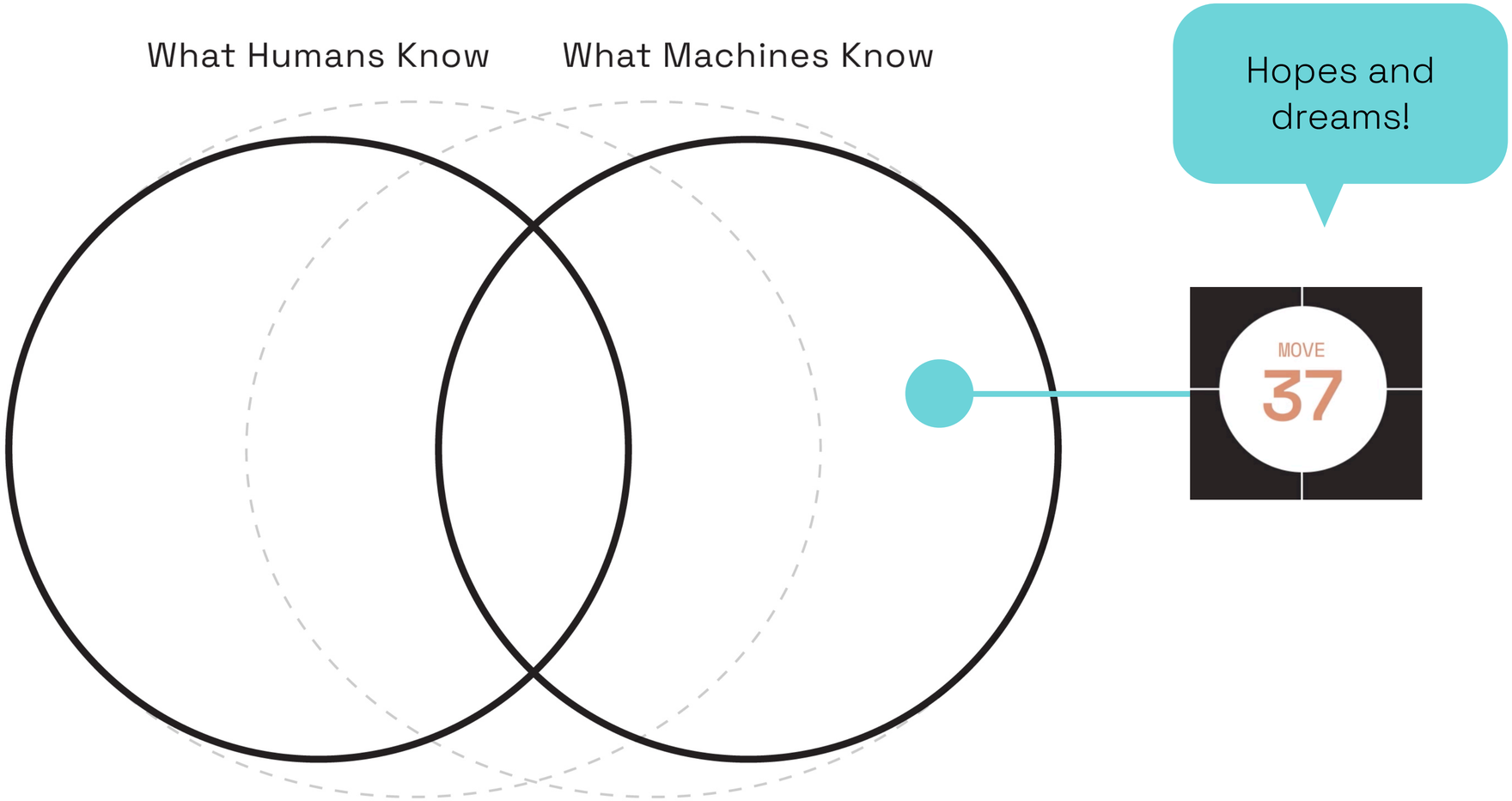
Expanding what we know: machines teaching humans new things.



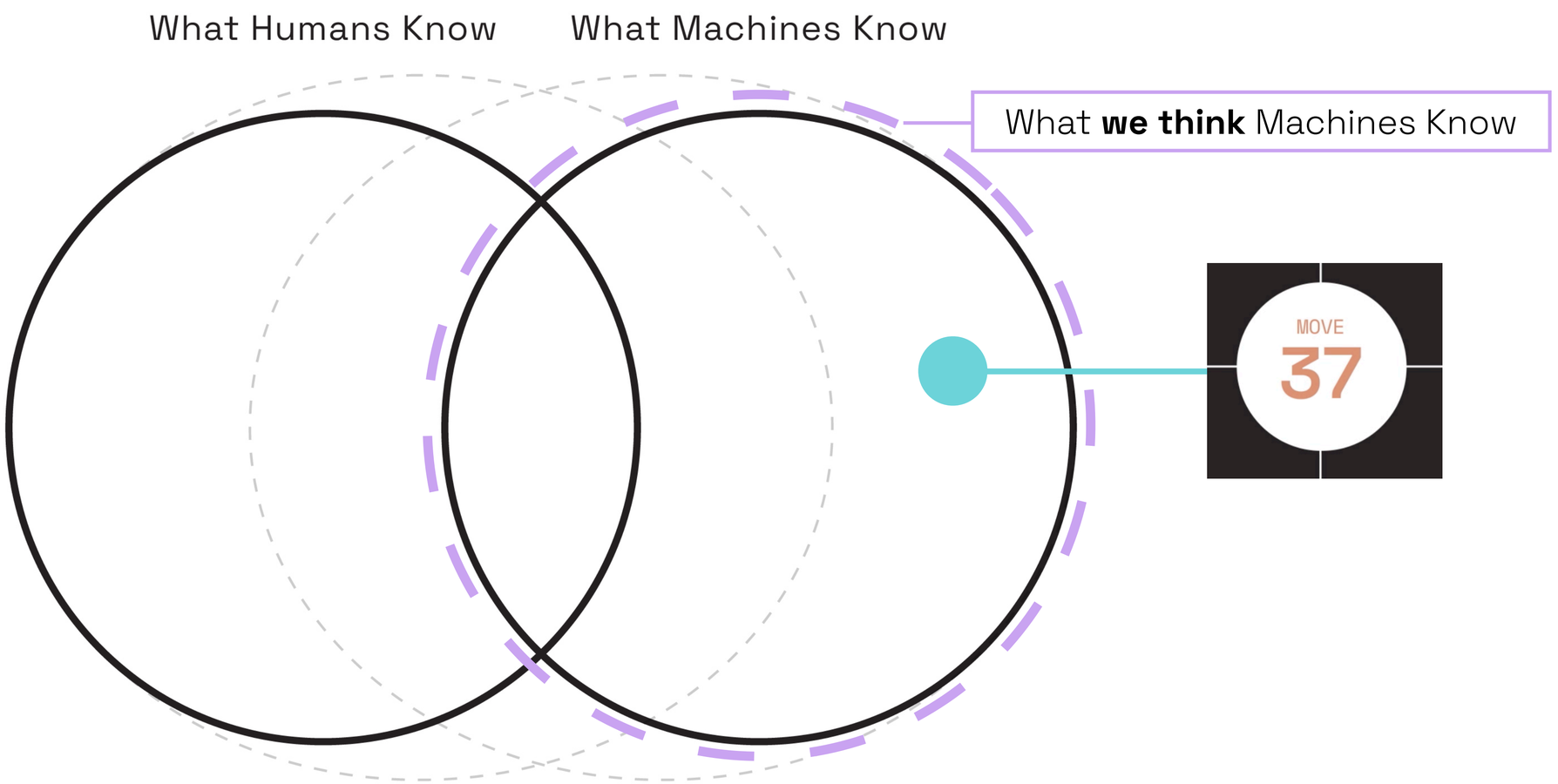

Hopes and dreams!
MOVE
37


Part 1 of
This talk
Hopes and
dreams!
MOVE
37

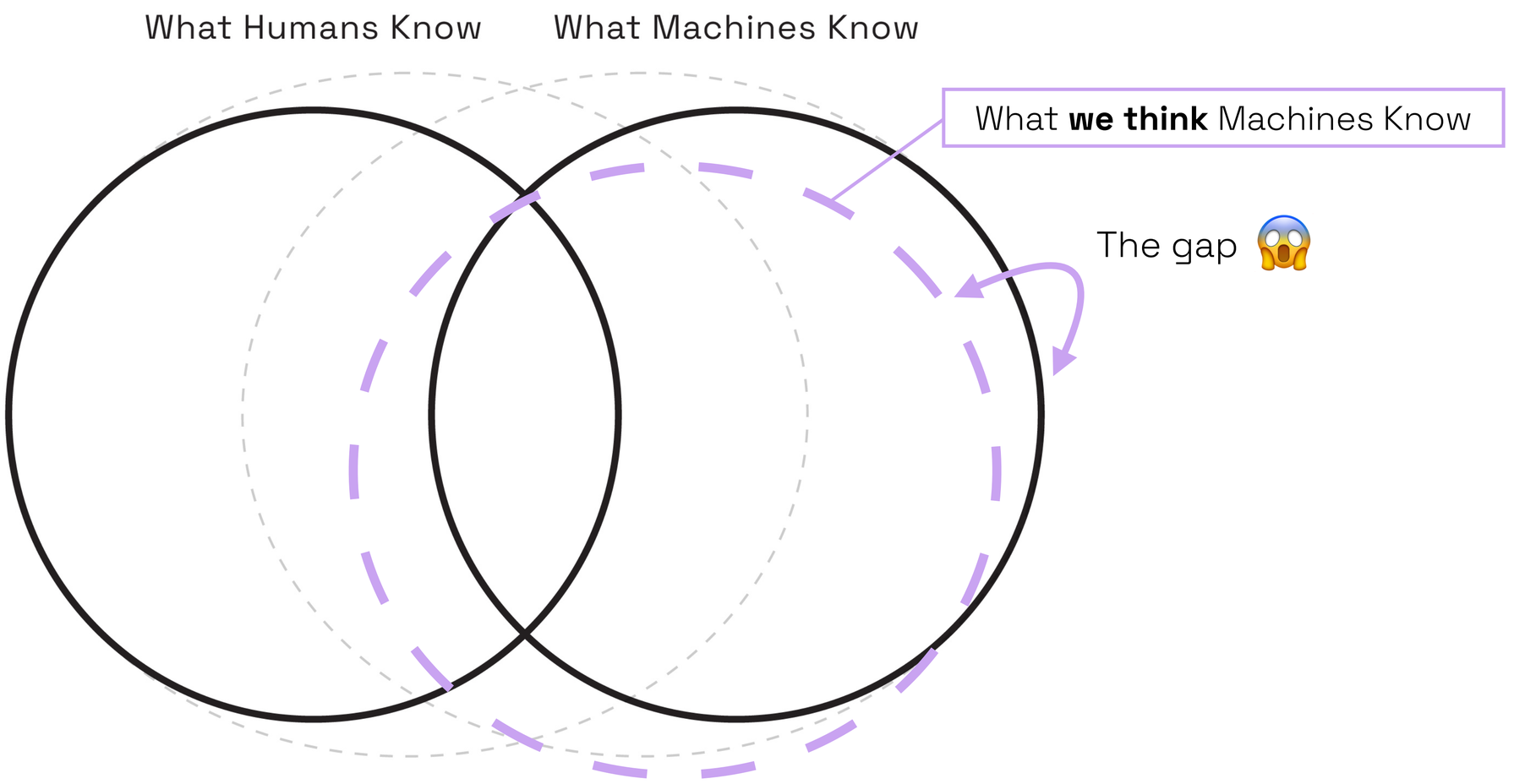

Hopes and dreams!
MOVE 37
RL
Part 2 of this talk
But all techniques are applicable to NLP.


What Humans Know
What Machines Know
Hopes and dreams!
MOVE
37

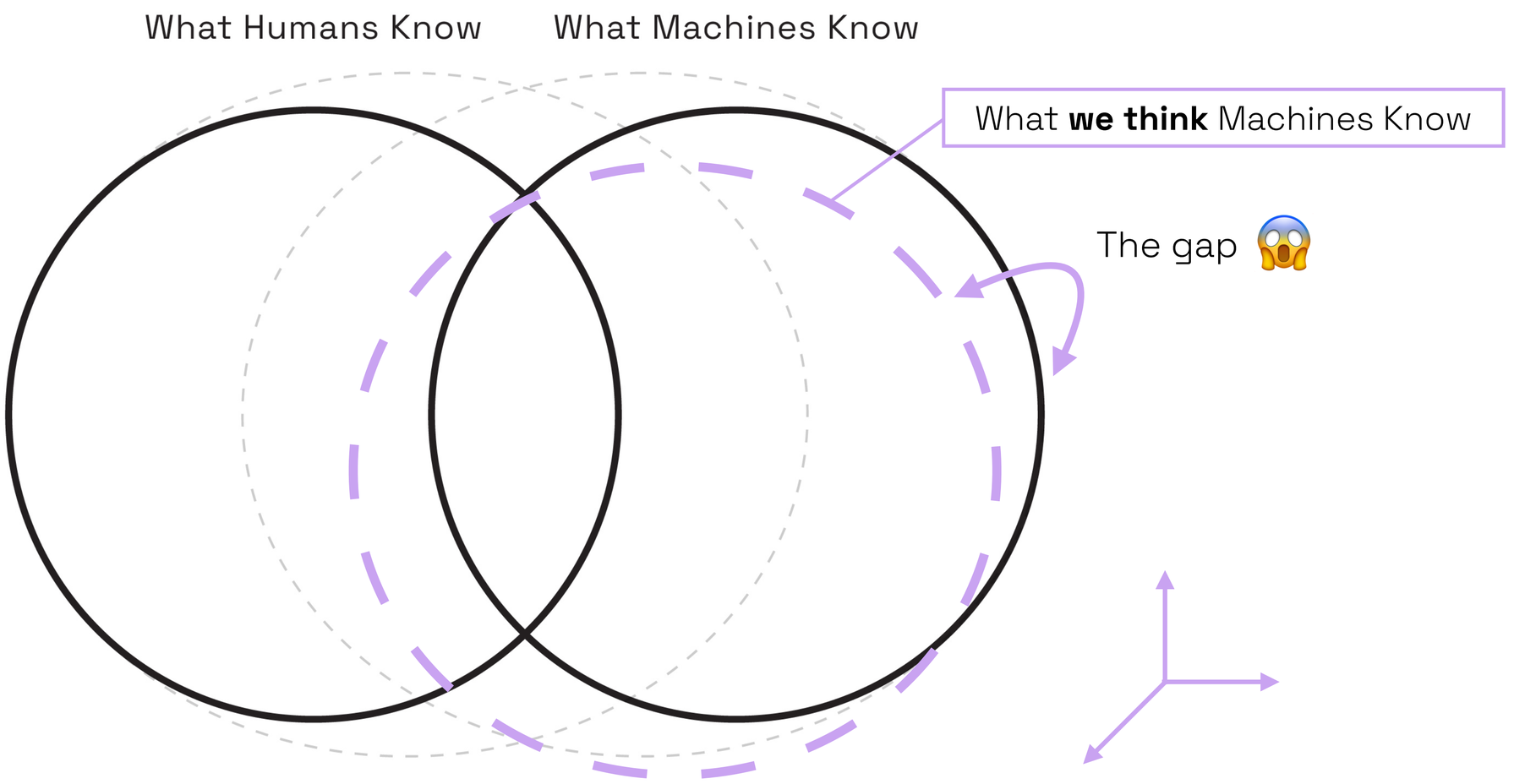
But first.. can we use the existing tools?



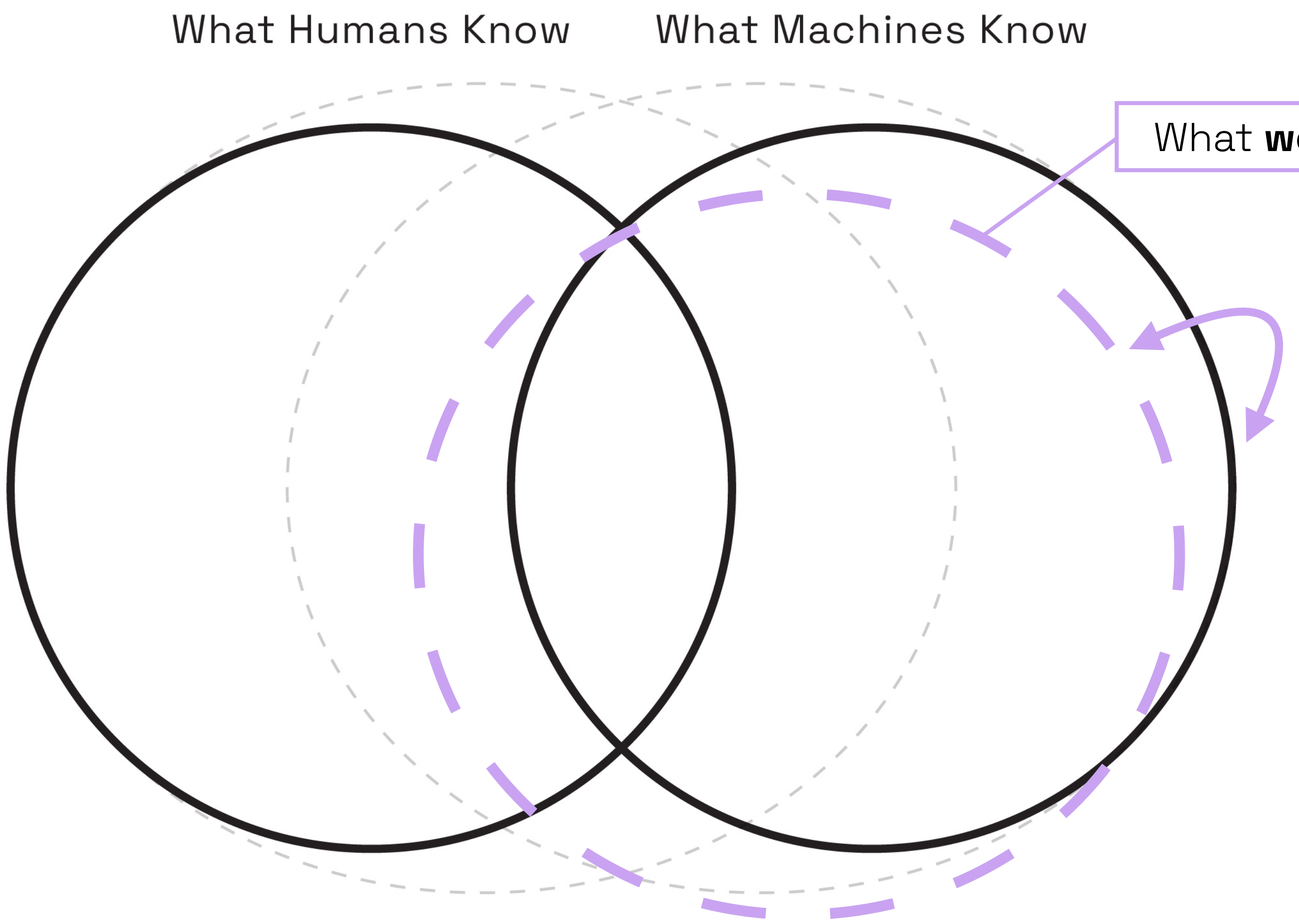
But first.. can we use the existing tools?



But first.. can we use the existing tools?

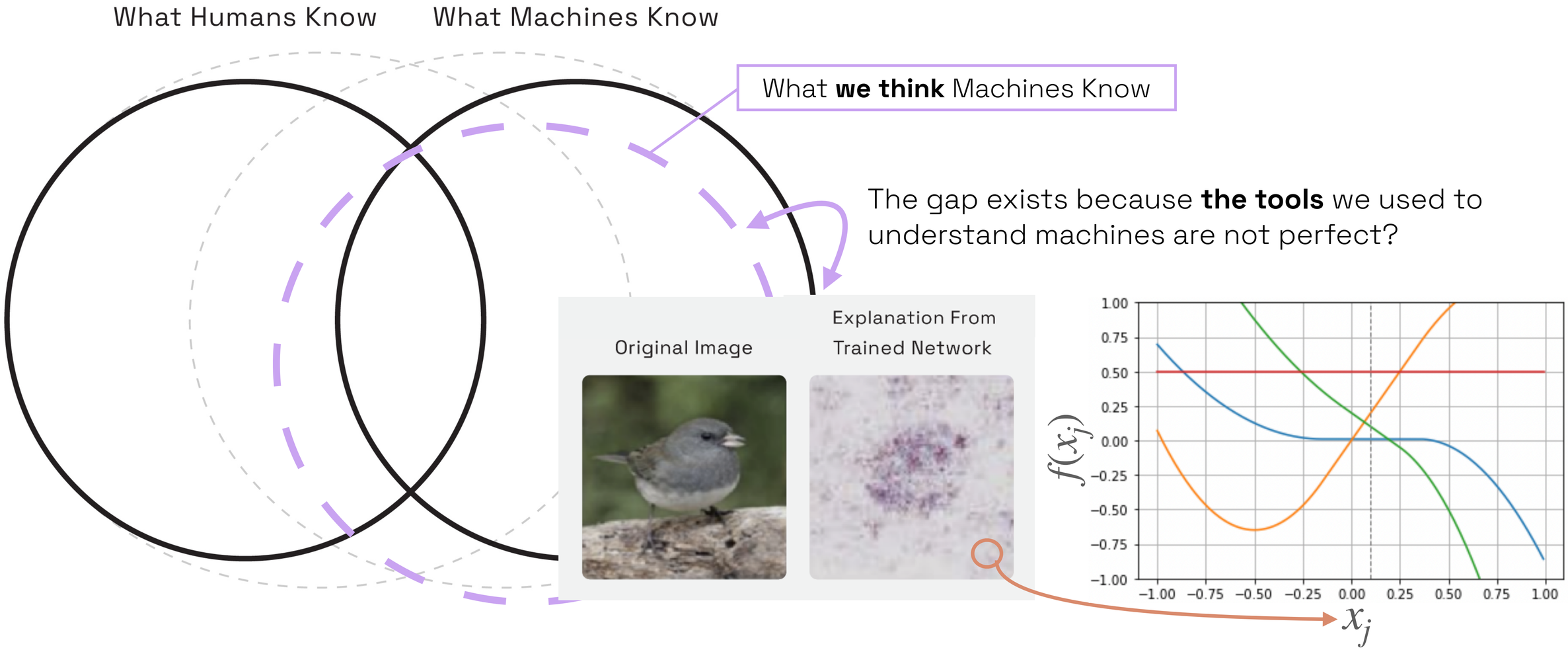


The gap between what machines know vs what we **think** machines know



The gap exists because **the tools** we used to understand machines are not perfect?

The gap between what machines know vs what we **think** machines know



The gap between what machines know vs what we **think** machines know

The gap exists because **the tools** we used to understand machines are not perfect?

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can’t understand them?

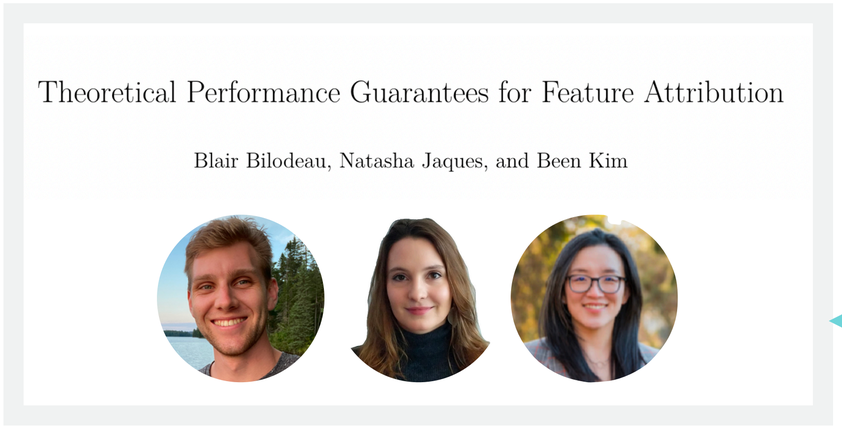
The gap between what machines know vs what we **think** machines know

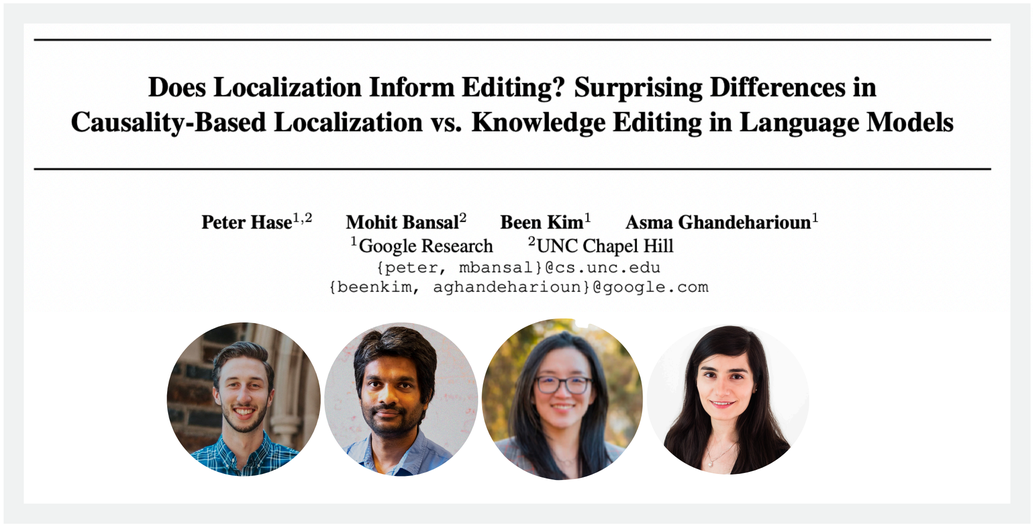
The gap exists because **the tools** we used to understand machines are not perfect?

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can’t understand them?

Gestalt phenomenon in Neural Networks
[K., Reif, Wattenberg, Bengio, Mozer, Comp. Brain & Behavior 2021]

The gap between what machines know vs what we **think** machines know



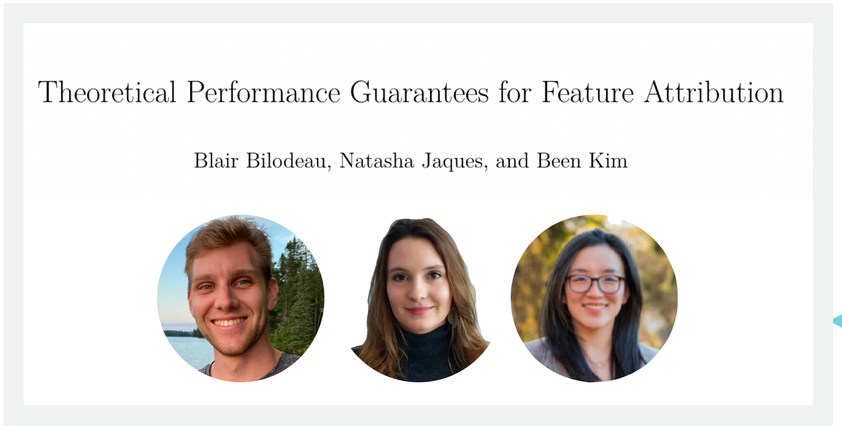


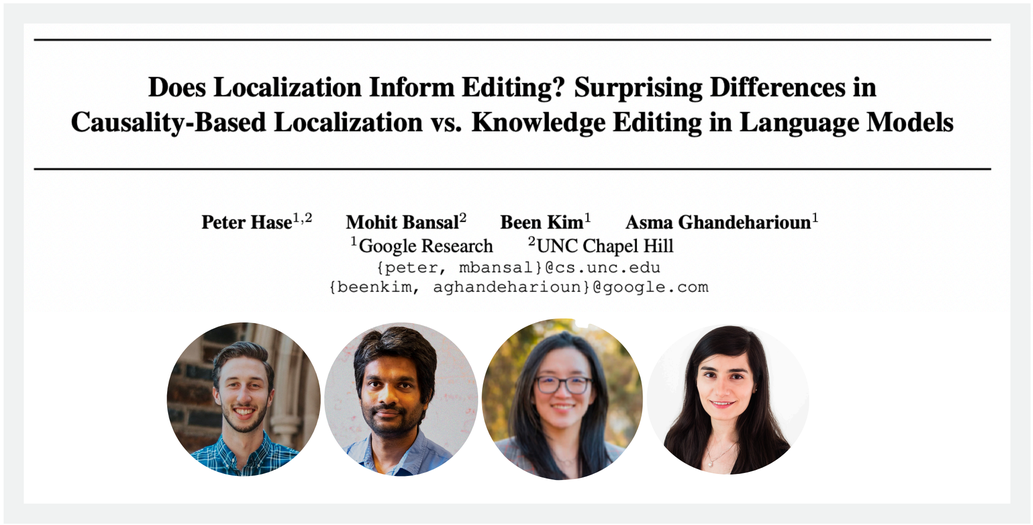
The gap exists because **the tools** we used to understand machines are not perfect?

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Gestalt phenomenon in Neural Networks
[K., Reif, Wattenberg, Bengio, Mozer, Comp. Brain & Behavior 2021]

The gap between what machines know vs what we **think** machines know





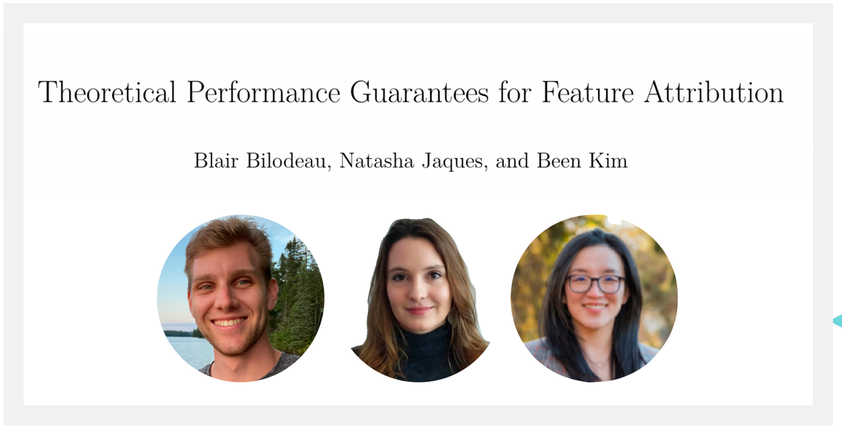
The gap exists because **the tools** we used to understand machines are not perfect?

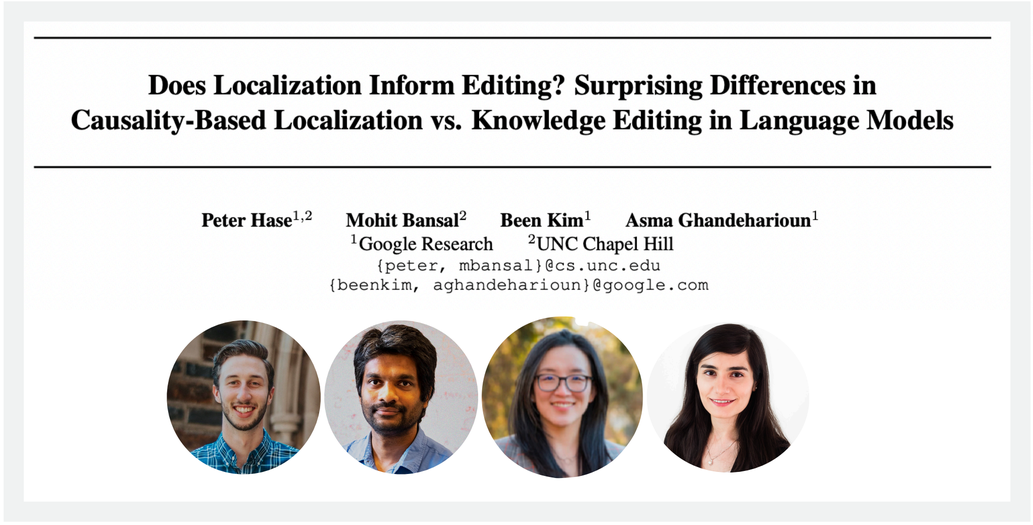
1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Debugging tests for model explanations
[Adabeyo, Muelly, Liccardi, K., Neurips 2020]
Post-hoc explanations may be ineffective detecting unknown spurious correlations
[Adabeyo, Muelly, Abelson, K., ICLR 2022]

Gestalt phenomenon in Neural Networks
[K., Reif, Wattenberg, Bengio, Mozer, Comp. Brain & Behavior 2021]

The gap between what machines know vs what we **think** machines know





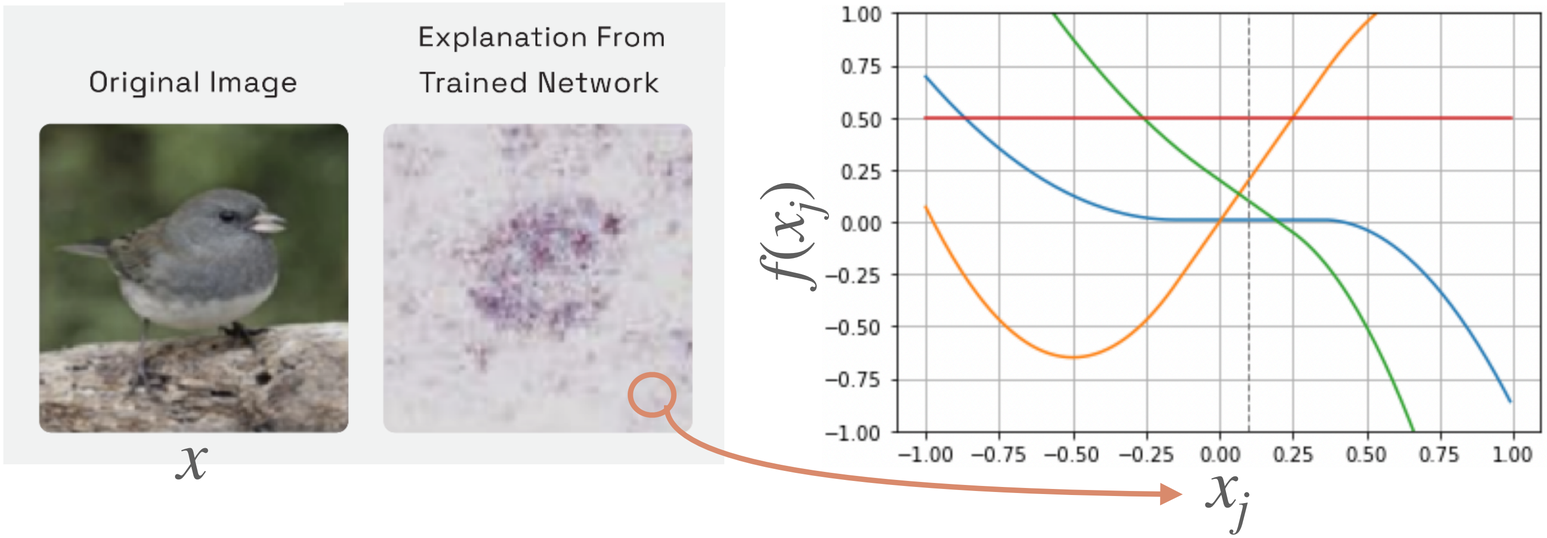
The gap exists because **the tools** we used to understand machines are not perfect?

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Debugging tests for model explanations
[Adabeyo, Muelly, Liccardi, K., Neurips 2020]
Post-hoc explanations may be ineffective detecting unknown spurious correlations
[Adabeyo, Muelly, Abelson, K., ICLR 2022]

Gestalt phenomenon in Neural Networks
[K., Reif, Wattenberg, Bengio, Mozer, Comp. Brain & Behavior 2021]

Earlier empirical work



Earlier empirical work

[NeurIPS 2018]

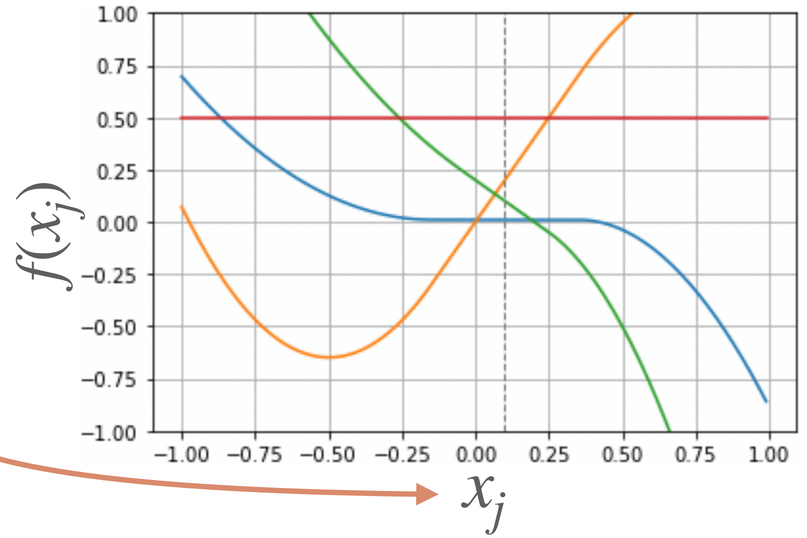
Sanity Checks for Saliency Maps

Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

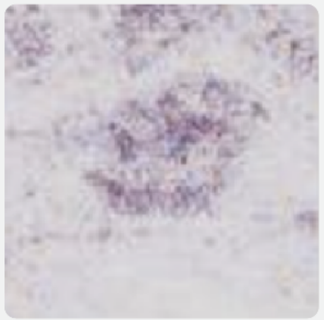
Original Image

x

Explanation From Trained Network



Explanation From Untrained Network



Explanations
from trained network
~ =
from untrained network ?!? 😱

Earlier empirical work

[NeurIPS 2018]

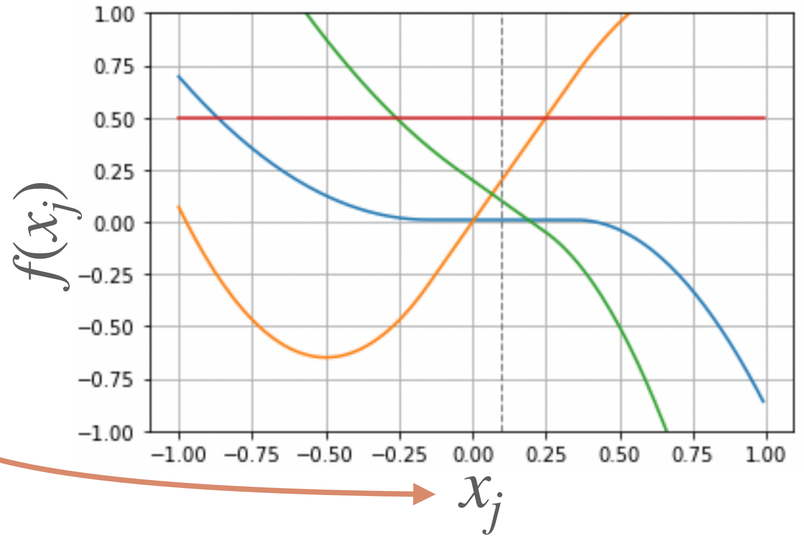
Sanity Checks for Saliency Maps

Julius Adebayo*, Justin Gilmer♯, Michael Muelly♯, Ian Goodfellow♯, Moritz Hardt♯†, Been Kim♯
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com
♯Google Brain
†University of California Berkeley

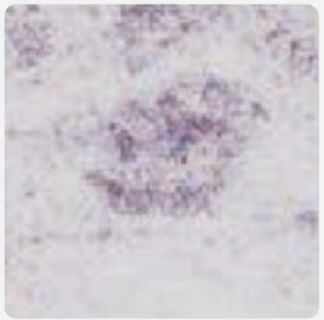
Original Image

x

Explanation From Trained Network



Explanation From Untrained Network



Explanations
from trained network
~ =
from untrained network ?!? 😱

But hey.. maybe it's still useful in practice!

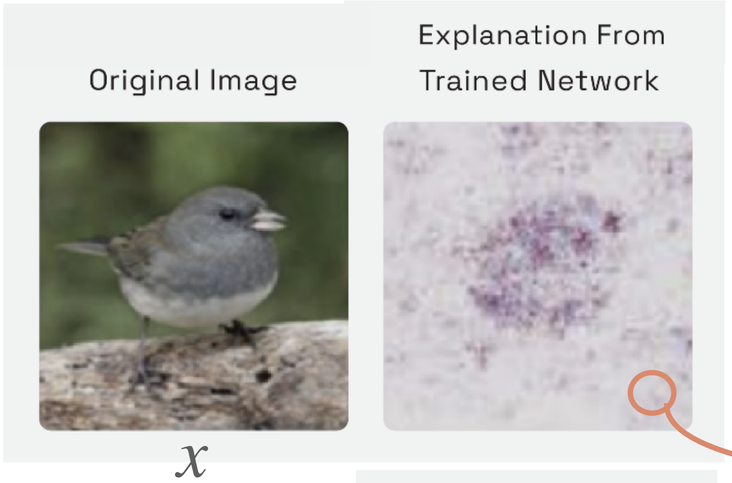
Earlier empirical work

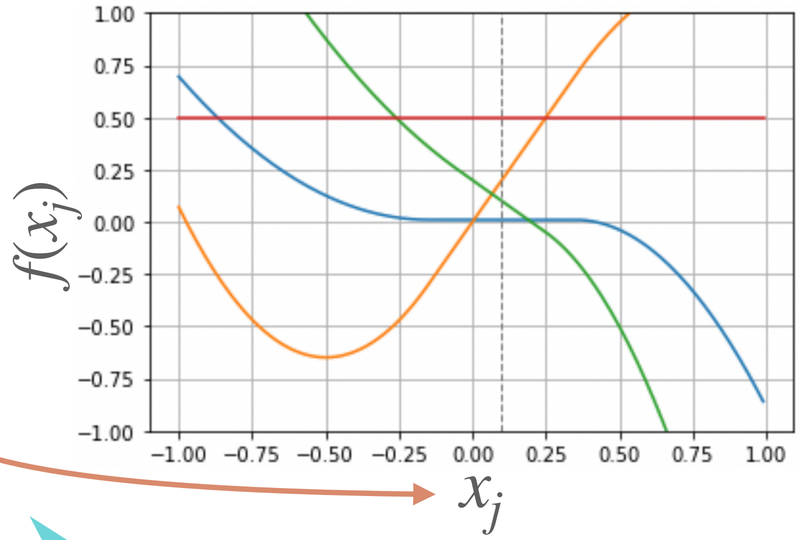
[ICLR 2022]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo* MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

Can explanations help people detect errors in practice?





Here is an input image and explanation. Would you deploy this model?

Bugs at each stage

- Labeling Errors
- Spurious Correlation
- Reinitialized Weights
- Unintentional frozen layers
- Out of distribution data
- Mismatch in preprocessing

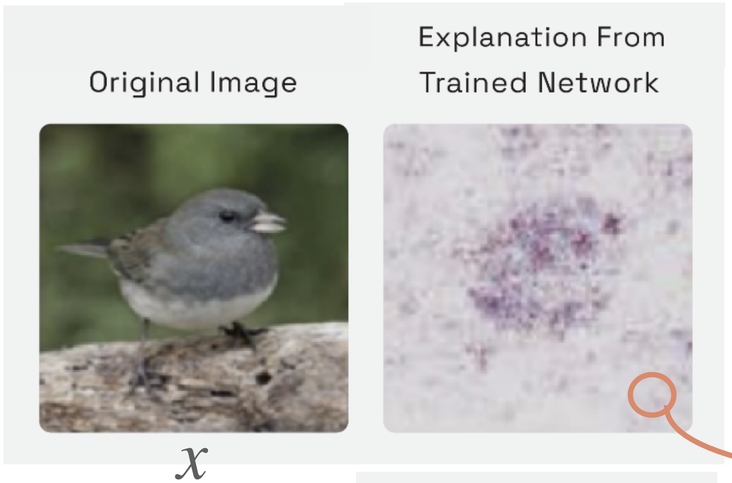
Earlier empirical work

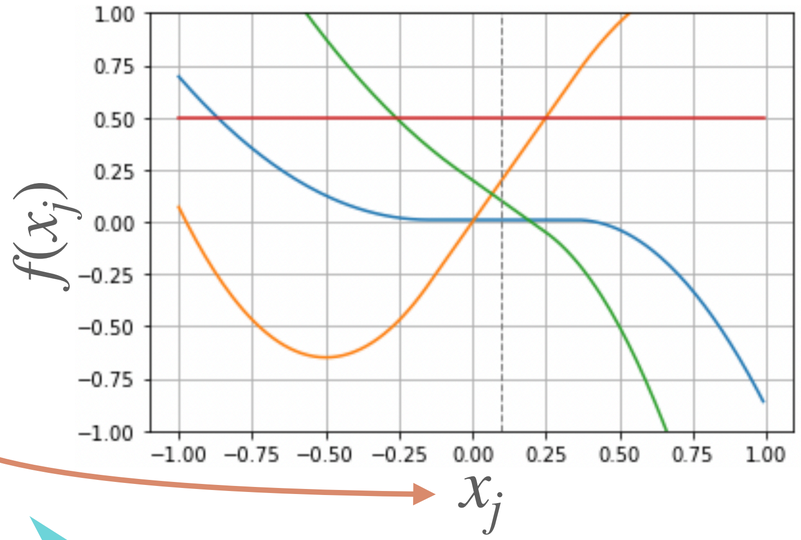
[ICLR 2022]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo* MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

Can explanations help people detect errors in practice?





Here is an input image and explanation. Would you deploy this model?

Bugs at each stage

No

- Labeling Errors
- Spurious Correlation

Maybe?

- Reinitialized Weights
- Unintentional frozen layers

No

- Out of distribution data
- Mismatch in preprocessing

No

No

😱

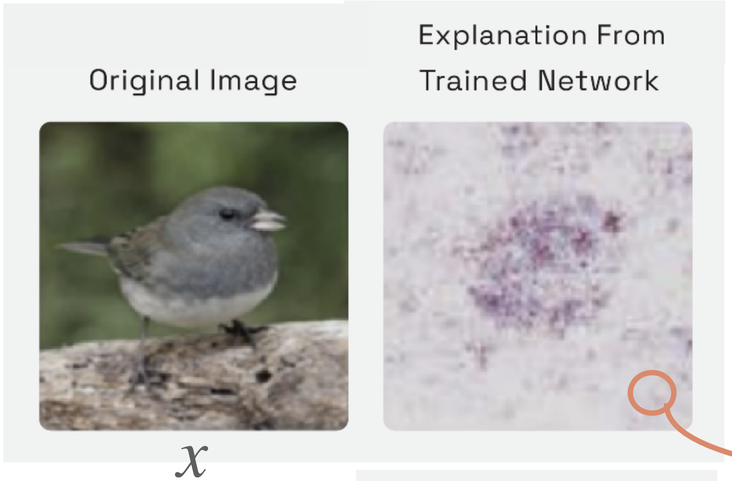
Earlier empirical work

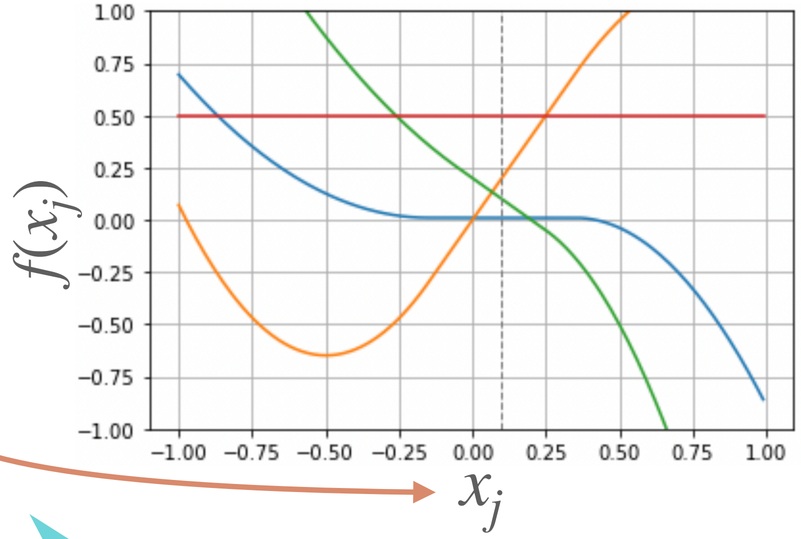
[ICLR 2022]

POST HOC EXPLANATIONS MAY BE INEFFECTIVE FOR DETECTING UNKNOWN SPURIOUS CORRELATION

Julius Adebayo* MIT CSAIL
Michael Muelly Stanford
Hal Abelson MIT CSAIL
Been Kim Google Research

Can explanations help people detect errors in practice?





Here is an input image and explanation. Would you deploy this model?

No

Bugs at each stage

- Labeling Errors
- Spurious Correlation

Maybe?

- Reinitialized Weights
- Unintentional frozen layers

No

- Out of distribution data
- Mismatch in preprocessing

No

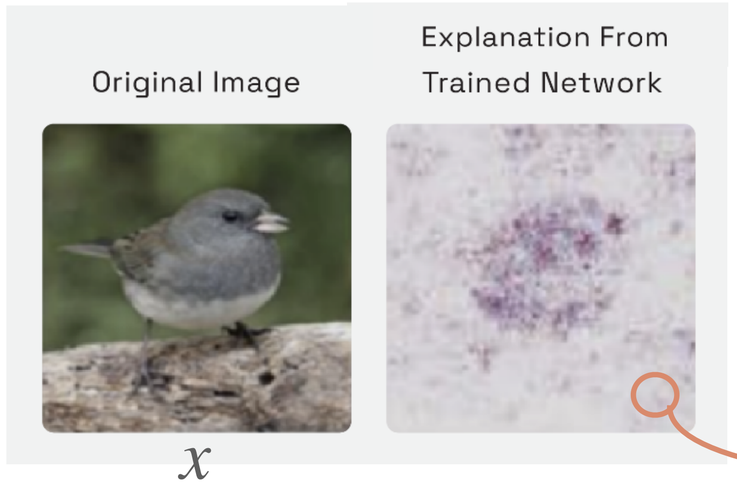
No

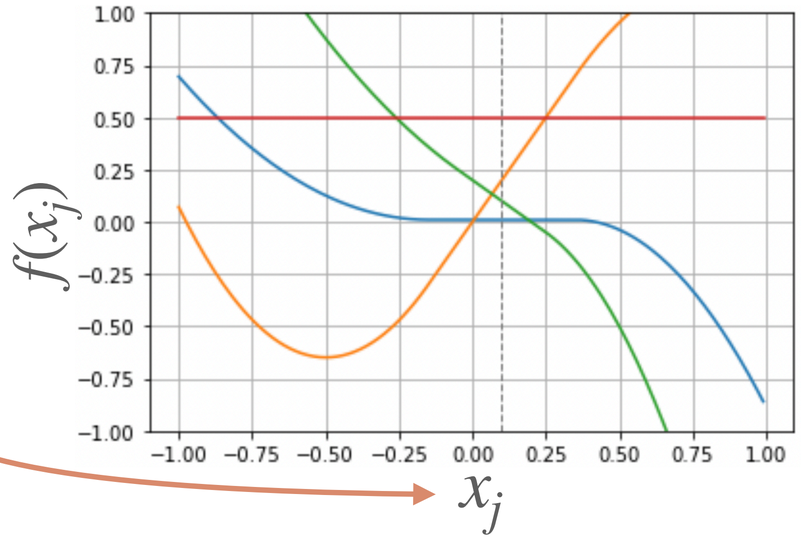
But hey.. there's no theoretic proof of all these! Maybe it's just one off!

TL;DR: The tools can be theoretically proven to be misaligned with our expectations.

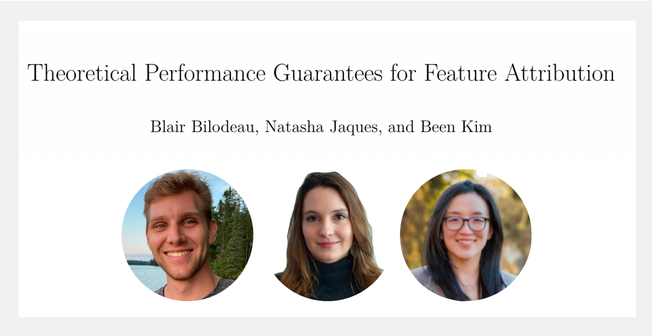
Theoretical Performance Guarantees for Feature Attribution

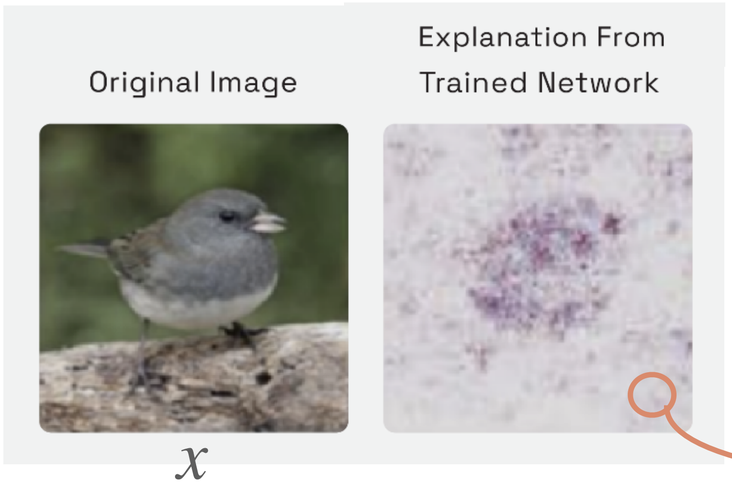
Blair Bilodeau, Natasha Jaques, and Been Kim

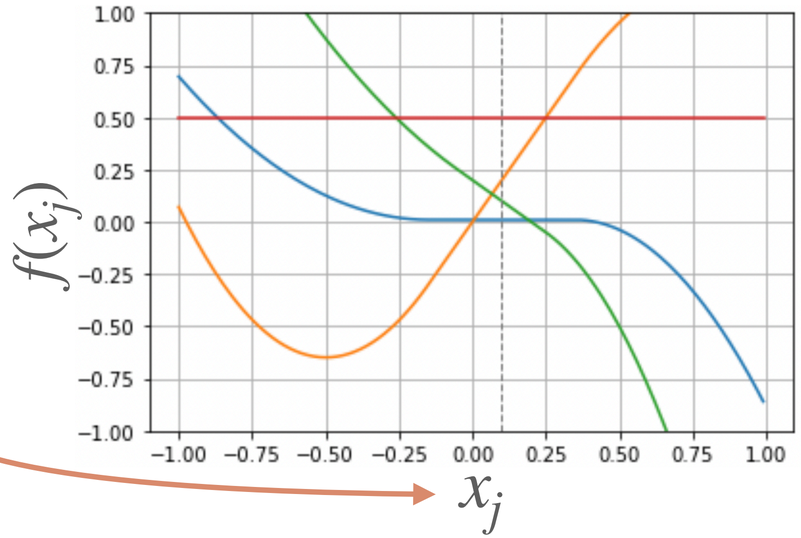




TL;DR: The tools can be theoretically proven to be misaligned with our expectations.





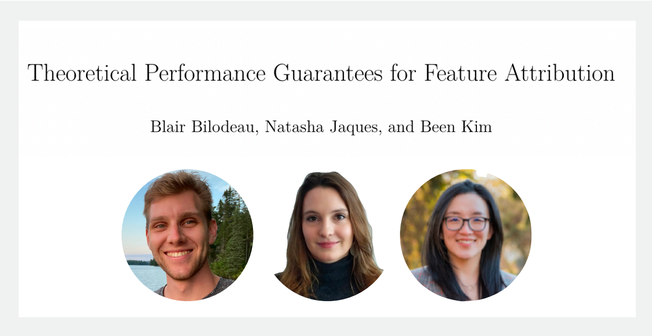


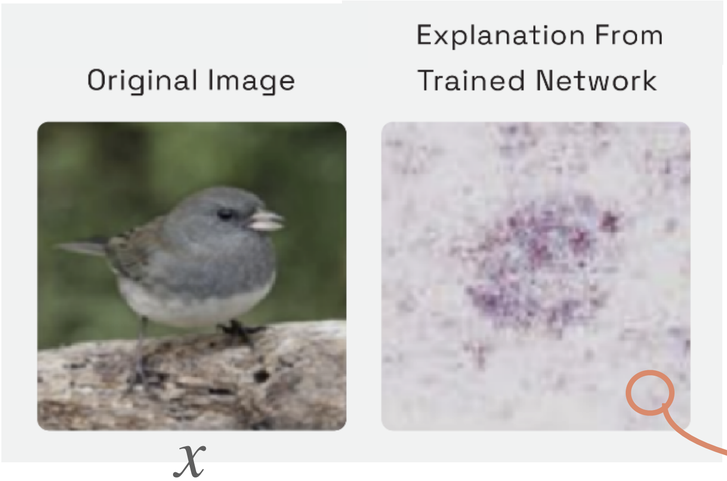
Expectations

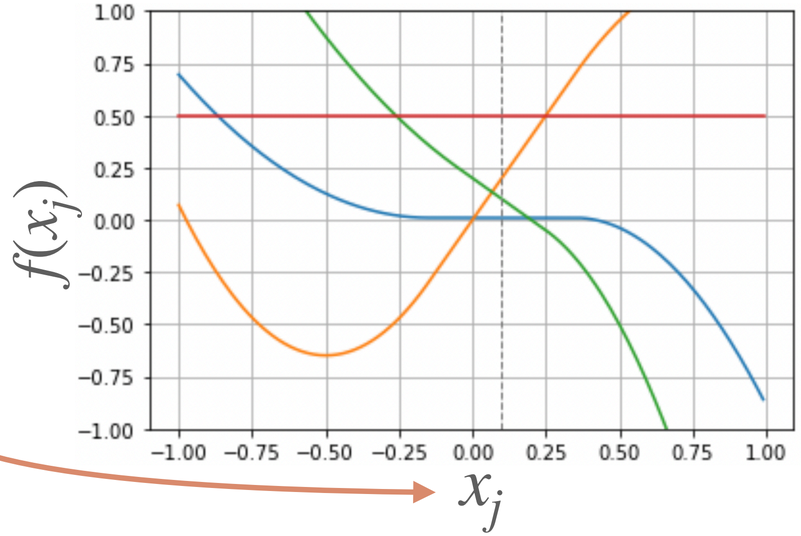
Integrated gradients paper [Sundararajan et al 17]
“Integrated gradients [...] **can be used for accounting the contributions** of each feature”

The tool assigned zero attribution to this pixel → This pixel is “unused” by the model f near x → f is insensitive to perturbations of this feature near x

TL;DR: The tools can be theoretically proven to be misaligned with our expectations.







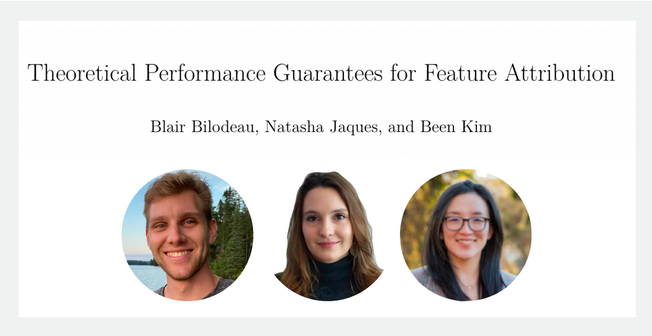
Expectations

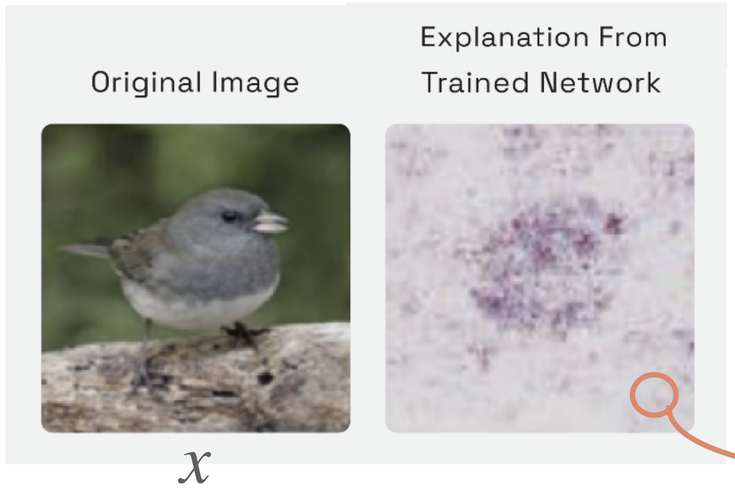
Integrated gradients paper [Sundararajan et al 17]
"Integrated gradients [...] **can be used for accounting the contributions** of each feature"

Evaluating Eligibility Criteria of OncologyTrials Using Real-World Data and AI [Liu et al. 21]
"Shapley values close to **zero [...] correspond to eligibility criteria that had no effect** on the hazard ratio of the overall survival."

The tool assigned zero attribution to this pixel → This pixel is "unused" by the model f near x → f is insensitive to perturbations of this feature near x

TL;DR: The tools can be theoretically proven to be misaligned with our expectations.

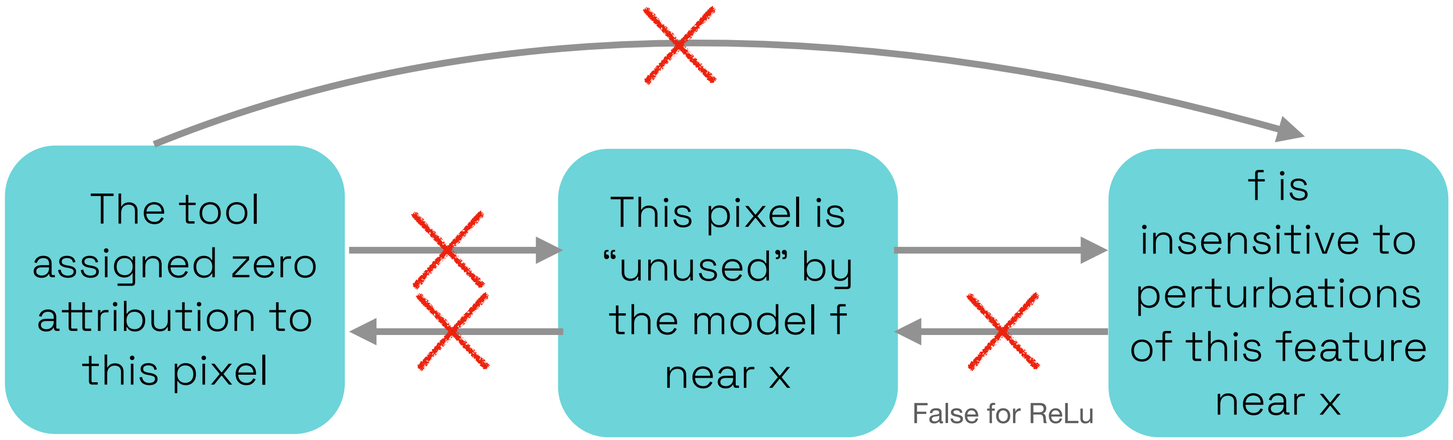




😱

TL;DR: Just because popular attribution methods* tell you there is “0” attribution to a feature, doesn’t mean you can conclude the model isn’t using the feature.

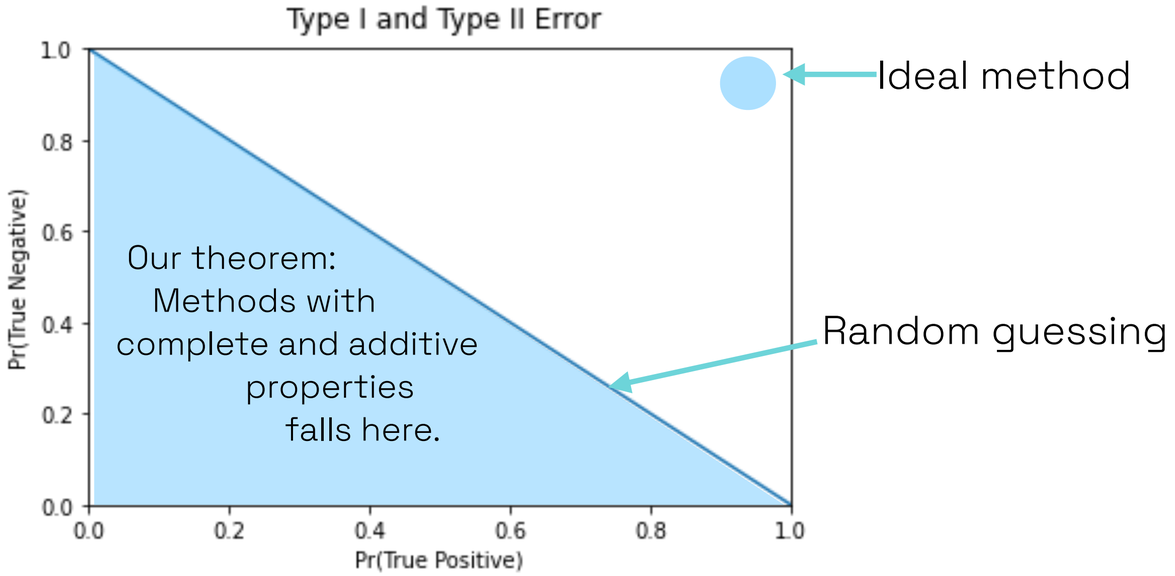
*attribution methods with completeness and additive properties (e.g., SHAP, integrated gradient)



Theoretical performance guarantees for feature attribution

- Main theorem sketch:
  - formulation: Interpreting attribution <> a hypothesis testing on the shape of the function (e.g., recourse, spurious correlation)
  - result: popular feature attribution methods (e.g., SHAP, IG) to conduct hypothesis testing about the model's behavior near a single data point (local explanation) implies:

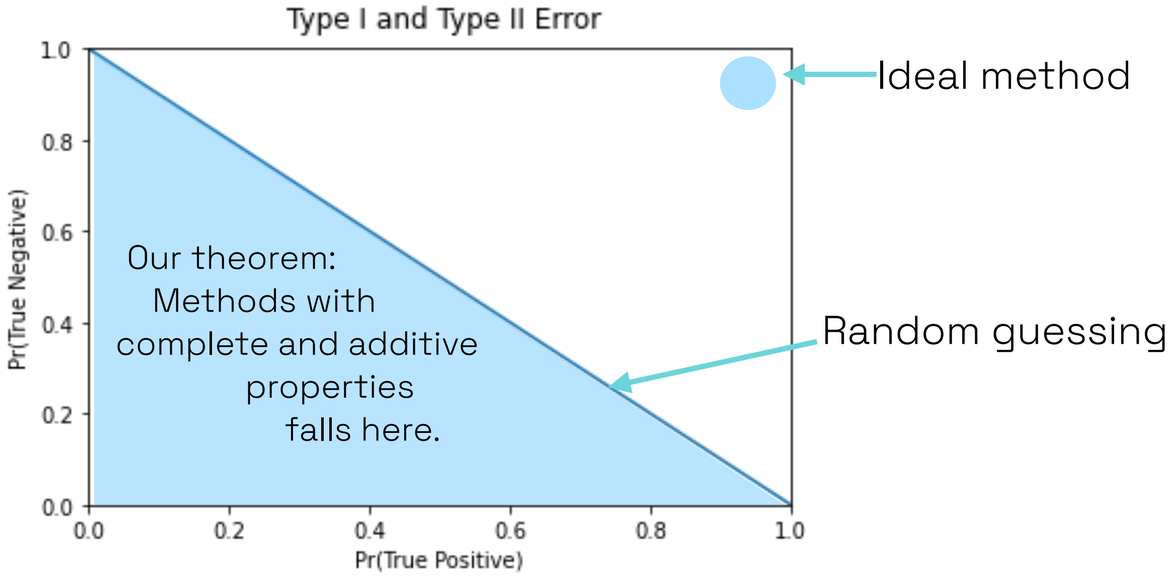
p(true positive) <= 1- p(true negative)



Theoretical performance guarantees for feature attribution

- Main theorem sketch:
  - formulation: Interpreting attribution <> a hypothesis testing on the shape of the function (e.g., recourse, spurious correlation)
  - result: popular feature attribution methods (e.g., SHAP, IG) to conduct hypothesis testing about the model's behavior near a single data point (local explanation) implies:

p(true positive) <= 1- p(true negative)



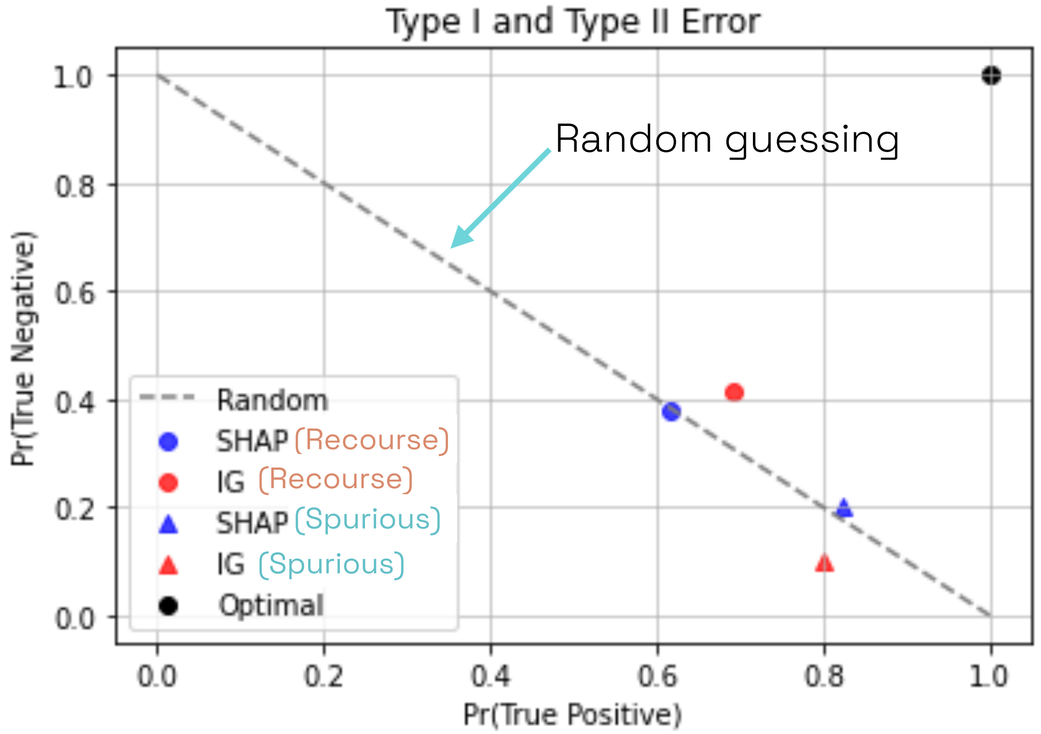
But wait, maybe they still "work" in practice?

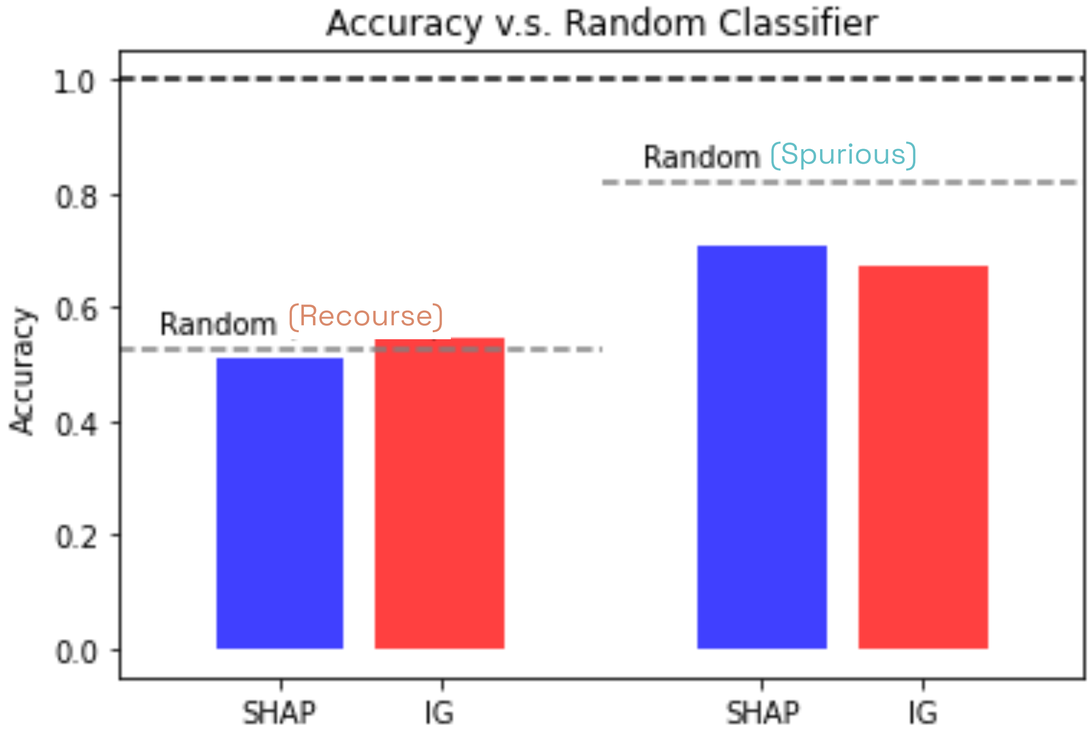
Do we observe this phenomenon in practice with real end-tasks?

Empirical validation of the theoretical performance guarantees

- How often this theory applies to practice? Often.
  - For two concrete downstream tasks
    - Recourse: which direction increases model's output probability?
    - Spurious features: which feature don't impact model output

Both reduces to the hypothesis testing



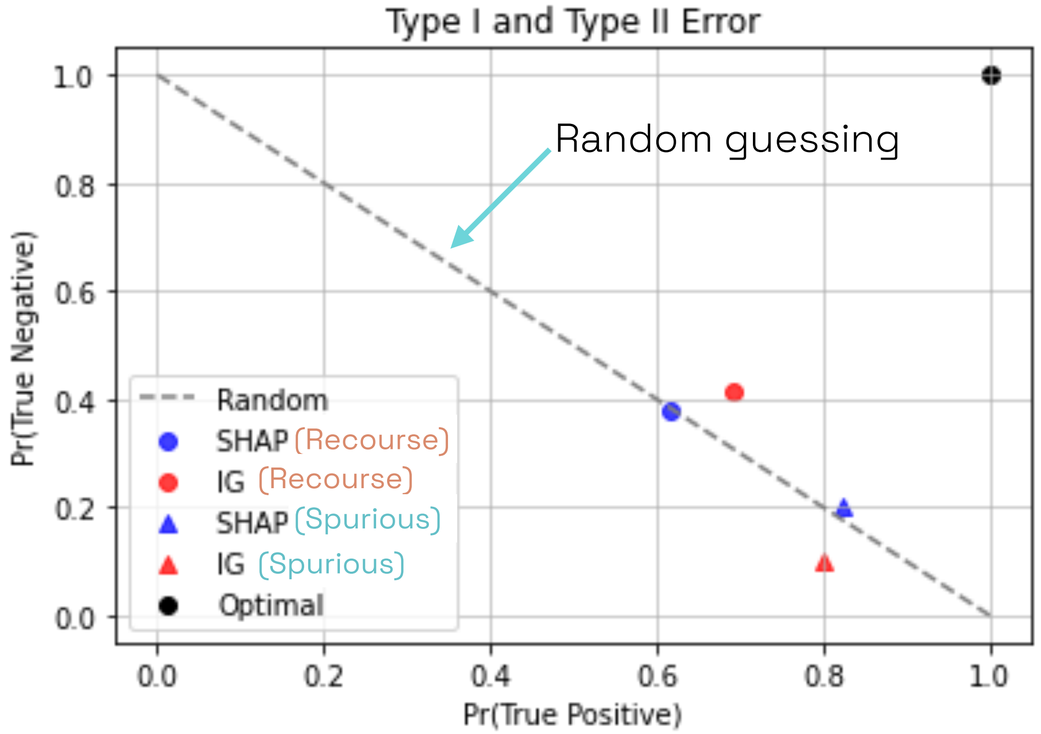


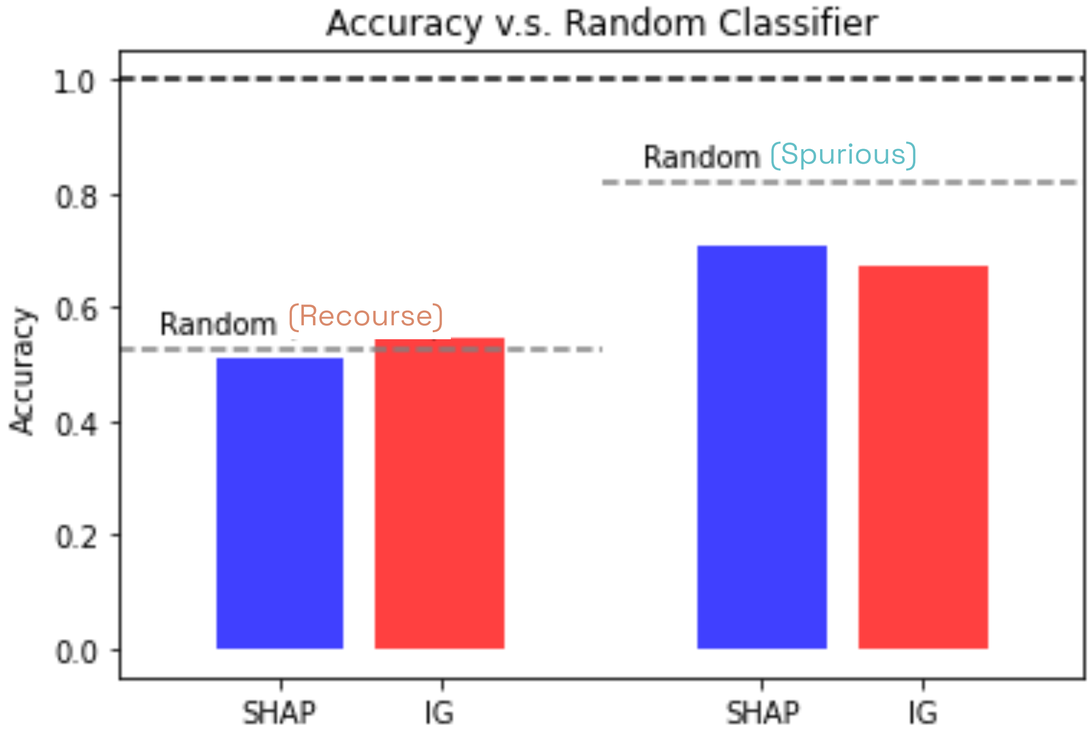
Empirical validation of the theoretical performance guarantees

- How often this theory applies to practice? Often.
  - For two concrete downstream tasks
    - Recourse: which direction increases model's output probability?
    - Spurious features: which feature don't impact model output

😱 Oh no!
Where do we go from here?!

→ Both reduces to the hypothesis testing





The answer might be simpler than developing more fancy methods: Use more samples.

- Inferring model behavior is impossible with current methods.
- One way forward: we can just brute force it.

  Question is: how many samples do we need to learn the shape of f?

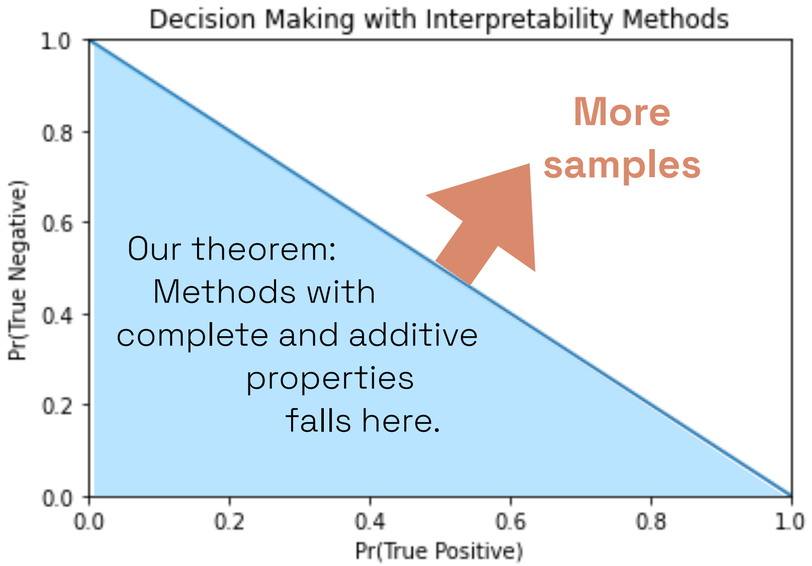
- Sampling complexity for spurious feature identification

p(true positive) <= 1- p(true negative) + $n\varepsilon^{-p}$

Number of features

Number of samples

differences output we wish to detect (resolution)



Summary and dirty laundry

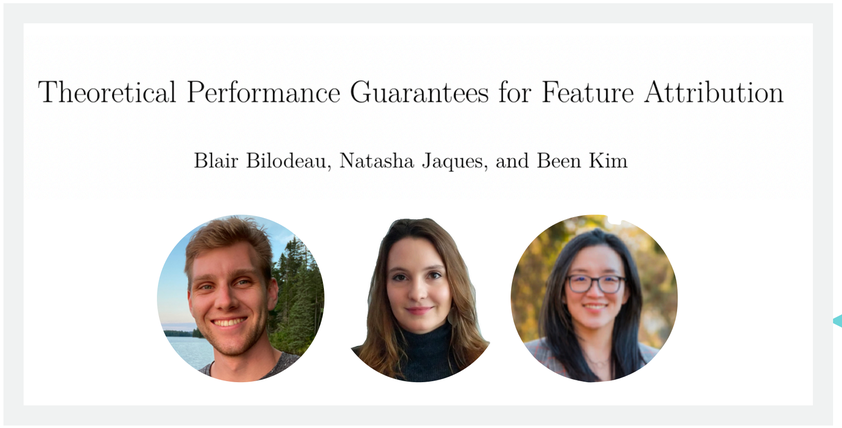
- Q. Can we infer model behavior with popular feature attribution methods?
- A. No

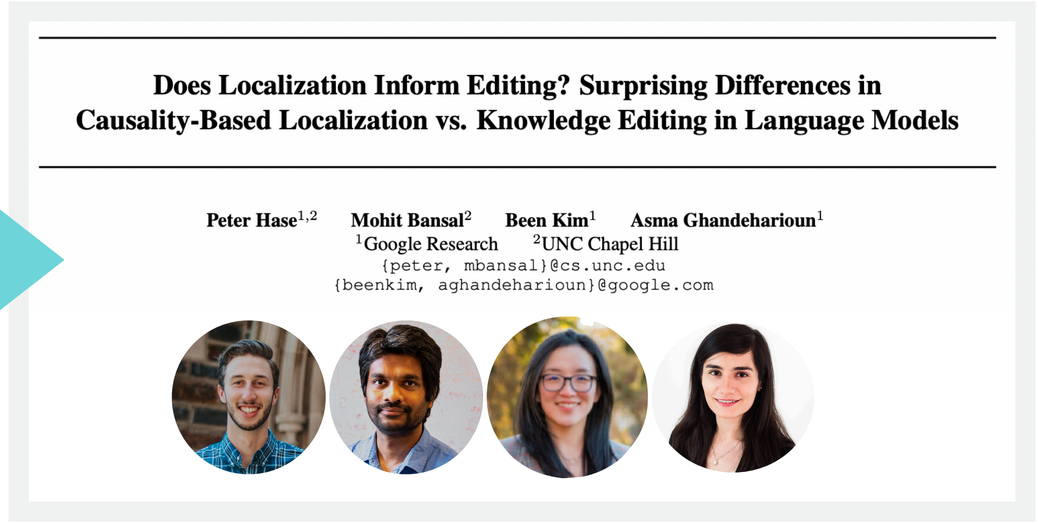
This holds both in theory and practice.

Future work

- Empirical validation of this phenomenon in bigger models.
- Identifying model-dependent sample complexity guarantees
- mathematically characterizing conditions (some narrower condition) under which the feature attribution methods meet our expectation [on-going work].

The gap between what machines know vs what we **think** machines know





The gap exists because **the tools** we used to understand machines have any of the following:

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Debugging tests for model explanations
[Adabeyo, Muelly, Liccardi, K., Neurips 2020]
Post-hoc explanations may be ineffective detecting unknown spurious correlations
[Adabeyo, Muelly, Abelson, K., ICLR 2022]

Gestalt phenomenon in Neural Networks
[K., Reif, Wattenberg, Bengio, Mozer, Comp. Brain & Behavior 2021]

One of those serendipity paper...

Let’s locate where ‘ethics’ knowledge is located in LLMs!
Maybe we can edit that knowledge to make ‘more ethically aware’ LLMs.

Let’s use an existing method to do this - ROME

Wait wait..

Something does’t line up.

TL;DR: model edit success is unrelated to location of factual information in models

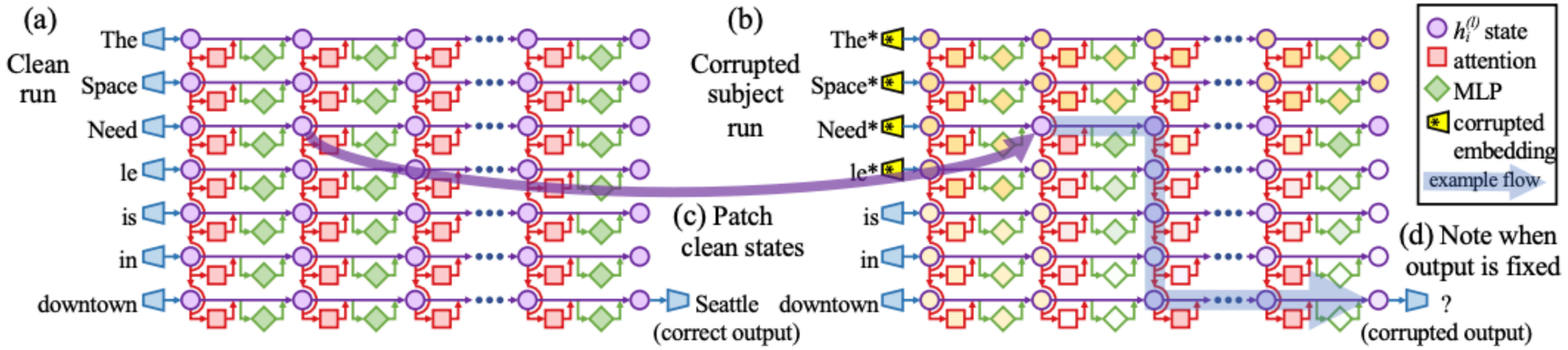
- Existing assumption:
  - Model editing is about changing what the model knows about.
  - We find where a fact is stored first then we edit it.

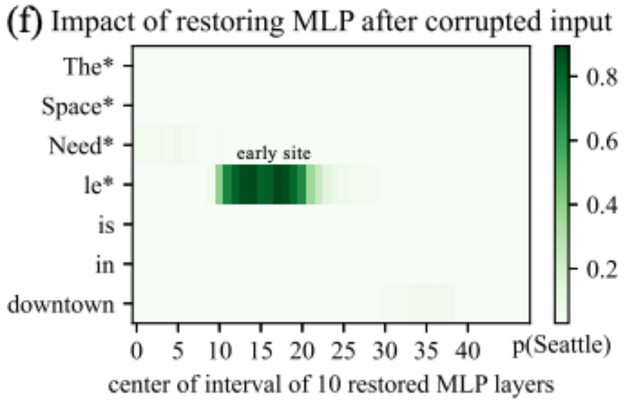
- TL;DR of our work: the assumption is not true.
  1. Substantial fraction of factual knowledge is stored **outside** of layers ‘identified’ as storing knowledge.
  2. The correlation between location vs editing is **near zero** (for ROME [Meng et al. 2022], MEMIT [Meng et al. 2022] and Adam-based fine-tuning).
  3. Our (unsuccessful) attempts to recover connection between location vs editing.

Method of interest: ROME and MEMIT [Meng et al. 2022 a, b]

- "Causal tracing" algorithm:
  1. Run the model on the input (s,r,o) e.g., (The space needle, located in, Seattle)
  2. Add noise to the embedding of s, then continue.
  3. Intervention: copy embeddings from 1 into 2, then continue.
  4. Calculate p(Seattle|noising, intervention)

Assumption:This tells us which layer to edit
(In reality ROME uses layer 6 for all edits, avg. causal tracing results)





Observation: many facts are stored outside of the layer 6.

- ROME only uses layer 6 to edit.
- But only 47% of facts peaks inside of blue region
  - Blue region: MEMIT is like ROME, except it uses multiple layers to edit.

So we wondered…

- Could we find better layer to edit instead of hard setting to layer 6?
- But first, a sanity check: more tracing effect <> better editing results?

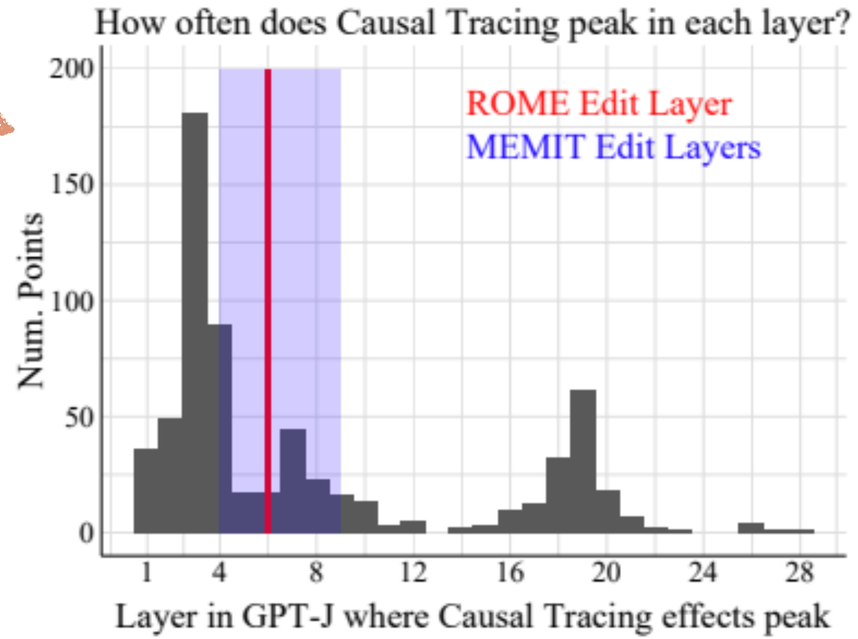


Figure 1: We visualize where a set of facts is stored in GPT-J, as localized by Causal Tracing. Model editing methods like ROME and MEMIT can successfully change knowledge in LMs by editing layers 4-9. But many facts appear to be stored outside of this range, e.g. at layers 1-3 and 16-20. What about these facts?

Does edit success correlated with tracing effect? - no 😱

Metrics:

- Edit success: Rewrite score: p(o-new) - successfully rewrite o to o-new?
- Tracing Effect (fraction restored):

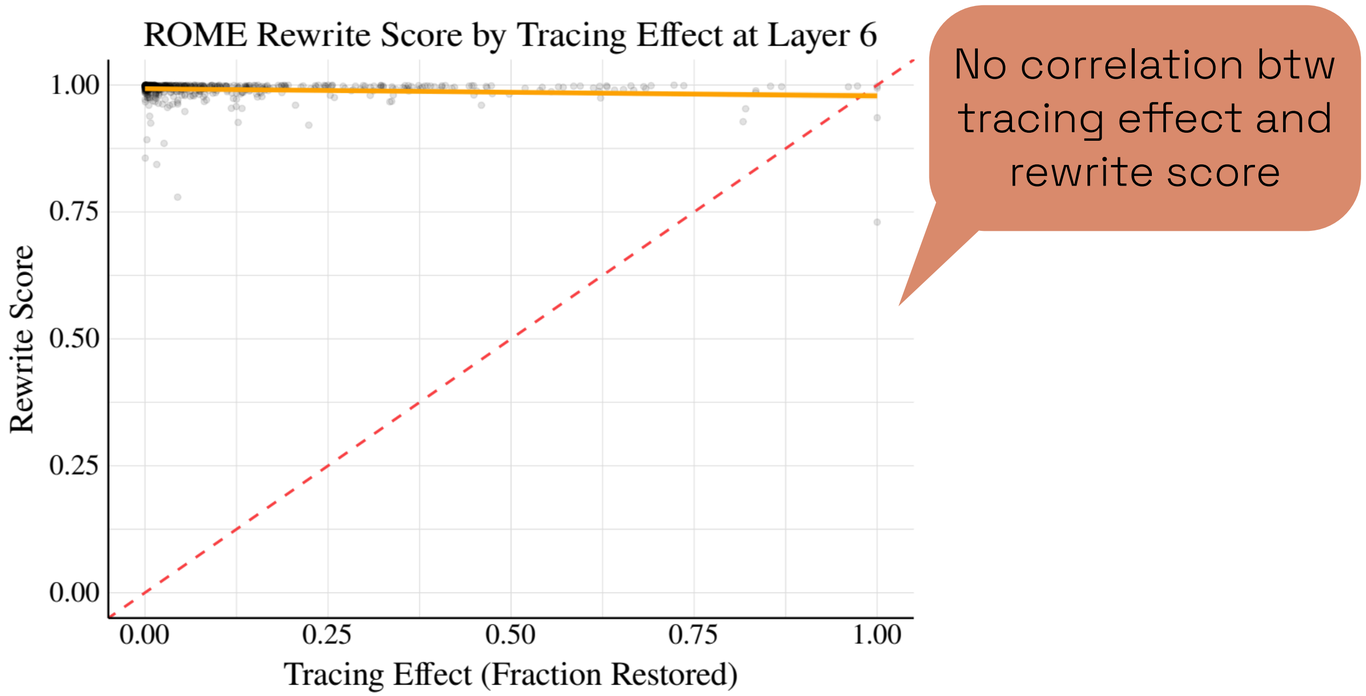
p(o-new | noising, intervention) - how much intervention restored the orig. value?

Does edit success correlated with intervention effect? - no 😱

Metrics:

- Edit success: Rewrite score: p(o-new) - successfully rewrite o to o-new?
- Tracing Effect (fraction restored):

p(o-new | noising, intervention) - how much intervention restored the orig. value?



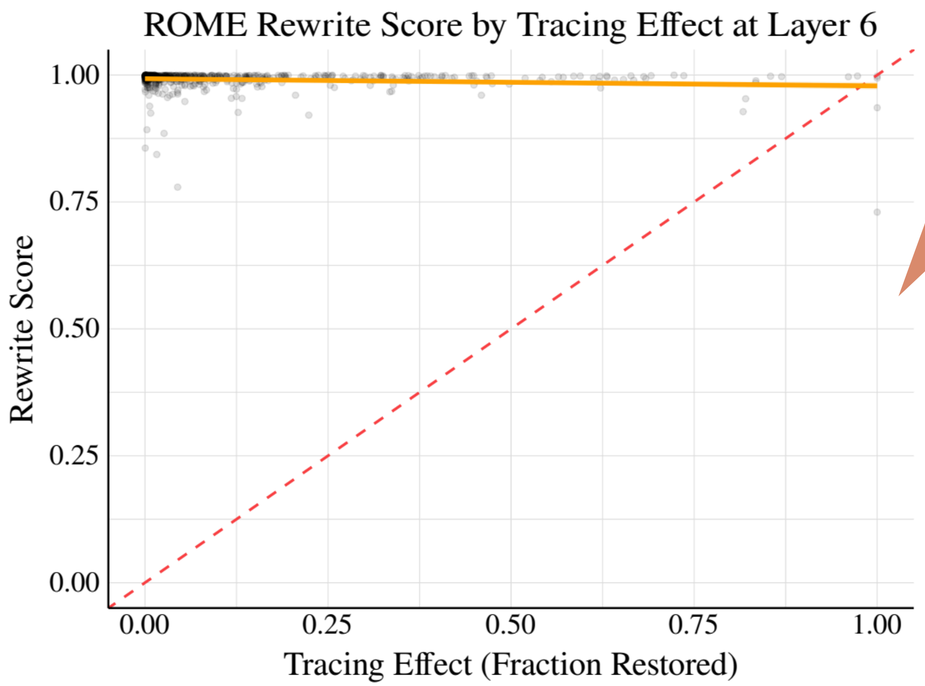
No correlation btw tracing effect and rewrite score

Does edit success correlated with intervention effect? - no 😱

Metrics:

- Edit success: Rewrite score: p(o-new) - successfully rewrite o to o-new?
- Tracing Effect (fraction restored):

p(o-new | noising, intervention) - how much intervention restored the orig. value?



No correlation btw tracing effect and rewrite score

how much the choice of layer vs tracing effect explains the variance in rewrite score: 94% layer

	R^2 Values		
Method	Layer	Tracing Effect	Both
ROME	0.947	0.016	0.948

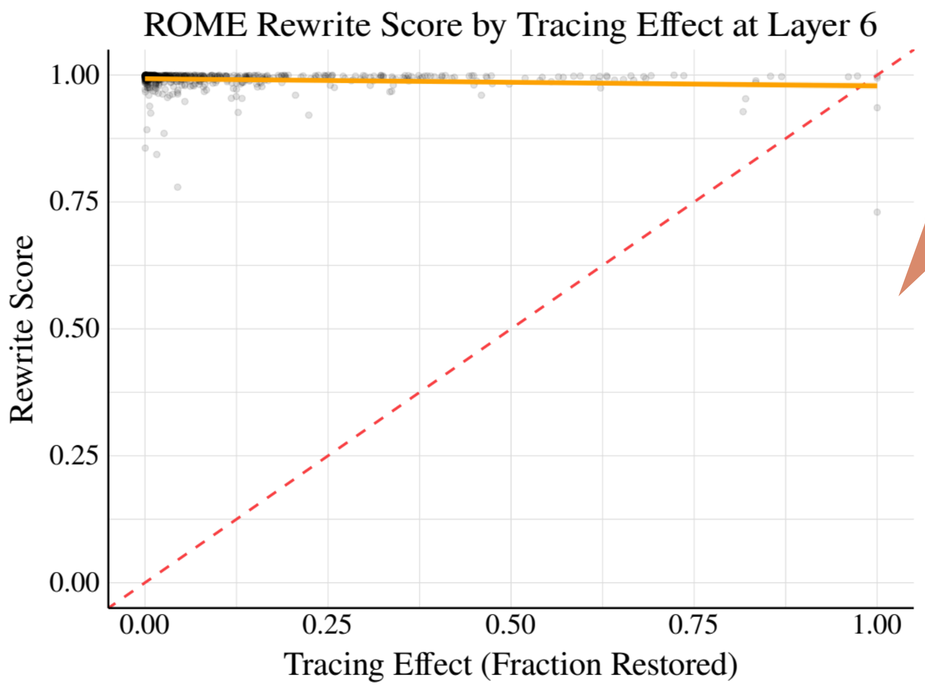
Table 1: R^2 values for predicting ROME edit success. Tracing effects explain essentially none of the variance in rewrite score, while the choice of edit layer is very important.

Does edit success correlated with intervention effect? - no 😱

Metrics:

- Edit success: Rewrite score: p(o-new) - successfully rewrite o to o-new?
- Tracing Effect (fraction restored):

p(o-new | noising, intervention) - how much intervention restored the orig. value?



No correlation btw tracing effect and rewrite score

how much the choice of layer vs tracing effect explains the variance in rewrite score: 94% layer

	R^2 Values		
Method	Layer	Tracing Effect	Both
ROME	0.947	0.016	0.948

Table 1: R^2 values for predicting ROME edit success. Tracing effects explain essentially none of the variance in rewrite score, while the choice of edit layer is very important.

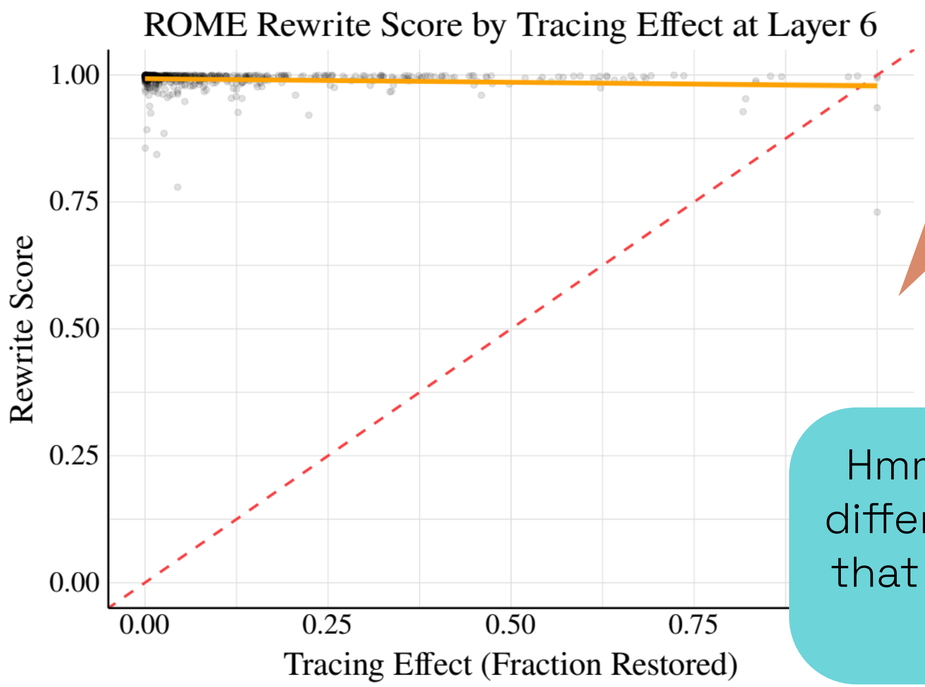
TL;DR: To edit a model, pick a good layer. Don't worry about localization.

Does edit success correlated with intervention effect? - no 😱

Metrics:

- Edit success: Rewrite score: p(o-new) - successfully rewrite o to o-new?
- Tracing Effect (fraction restored):

p(o-new | noising, intervention) - how much intervention restored the orig. value?



No correlation btw tracing effect and rewrite score

how much the choice of layer vs tracing effect explains the variance in rewrite score: 94% layer

	R^2 Values		
Method	Layer	Tracing Effect	Both
ROME	0.947	0.016	0.948

Table 1: R^2 values for predicting ROME edit success. Tracing effects explain essentially none of the variance in rewrite score, while the choice of edit layer is very important.

Hmm.. maybe there are different kinds of editing that does correlate with localization?

TL;DR: To edit a model, pick a good layer. Don't worry about localization.

Are there different kinds of editing where localization is important? - no

Editing Problem Variants	Input Prompt	Objective
Error Injection	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(\text{Sweden}\|\text{Input})$
Tracing Reversal	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(o_{\text{noise}}\|\text{Input})$
Fact Erasure	Autonomous University of Madrid, which is located in _____	$\arg\min_\theta p_\theta(\text{Spain}\|\text{Input})$
Fact Amplification	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(\text{Spain}\|\text{Input})$
Fact Forcing	*Autonomous University of Madrid*, which is located in _____ (Add noise to subject)	$\arg\max_\theta p_\theta(\text{Spain}\|\text{Noisy Input})$

Introducing new kinds of editing

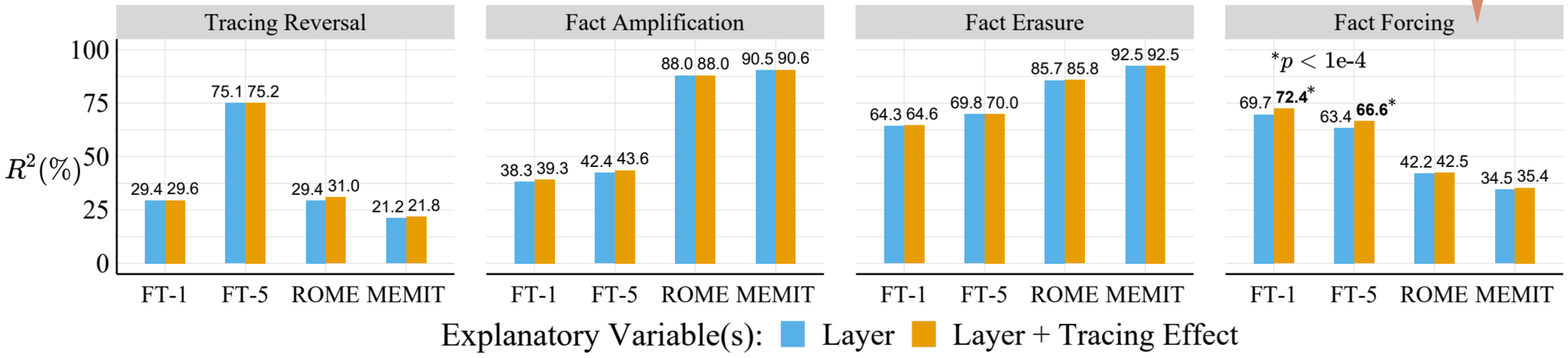
Are there different kinds of editing where localization is important? - no

Editing Problem Variants	Input Prompt	Objective
Error Injection	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(\text{Sweden}\|\text{Input})$
Tracing Reversal	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(o_{\text{noise}}\|\text{Input})$
Fact Erasure	Autonomous University of Madrid, which is located in _____	$\arg\min_\theta p_\theta(\text{Spain}\|\text{Input})$
Fact Amplification	Autonomous University of Madrid, which is located in _____	$\arg\max_\theta p_\theta(\text{Spain}\|\text{Input})$
Fact Forcing	*Autonomous University of Madrid* (Add noise to subject), which is located in _____	$\arg\max_\theta p_\theta(\text{Spain}\|\text{Noisy Input})$

Introducing new kinds of editing

Fact forcing may look very promising but choice of edit layer is still by far the more important factor

Tracing effects are very weakly predictive of edit success



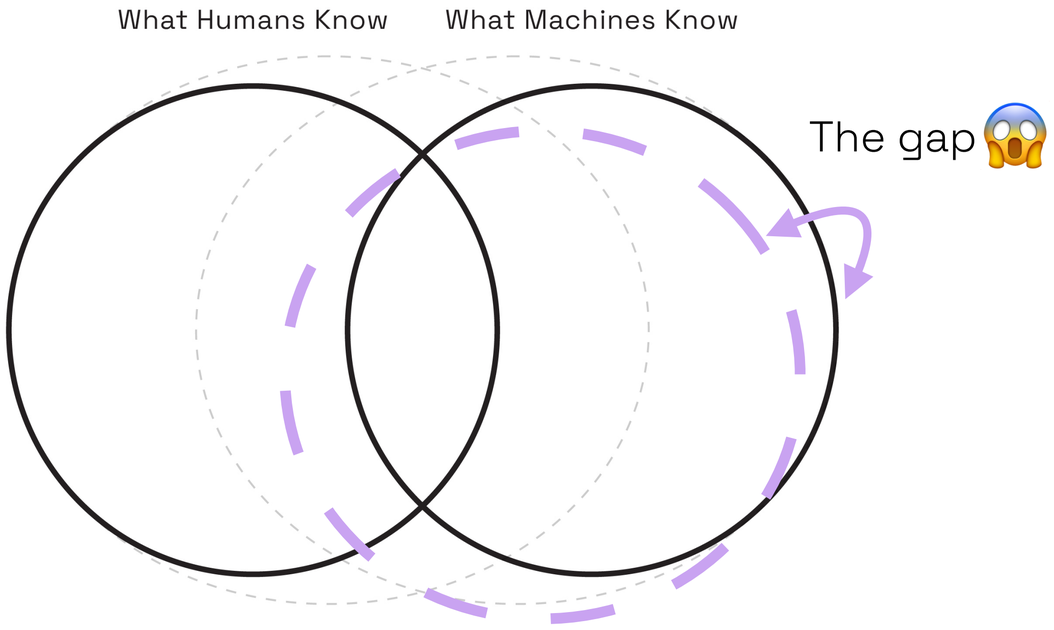
Where does it leave us?

- Q. Localize helps editing?
- A. no. There is no relationship between localization and editing success (for current localization methods, GPT-j + CounterFact data)
- Q. Are there any other editing that correlate better with localization?
- A. no.
- Causal tracing = factual information is carried in presentations in Transformer's forward pass, and that only. != where is best to intervene to change the factual information.

Thoughts:

- Causal tracing revealed the role that early-to-mid-range MLP representations at the last subject token index play in factual association. -> useful
- Important NOT to (1) validate the results of localization via editing or (2) motivate the editing method via localization.

Bridging the gap between what machines know vs. what we think they know.



Good artists steal.
Good researchers doubt.

The gap exists because **the tools** we used to understand machines have any of the following:

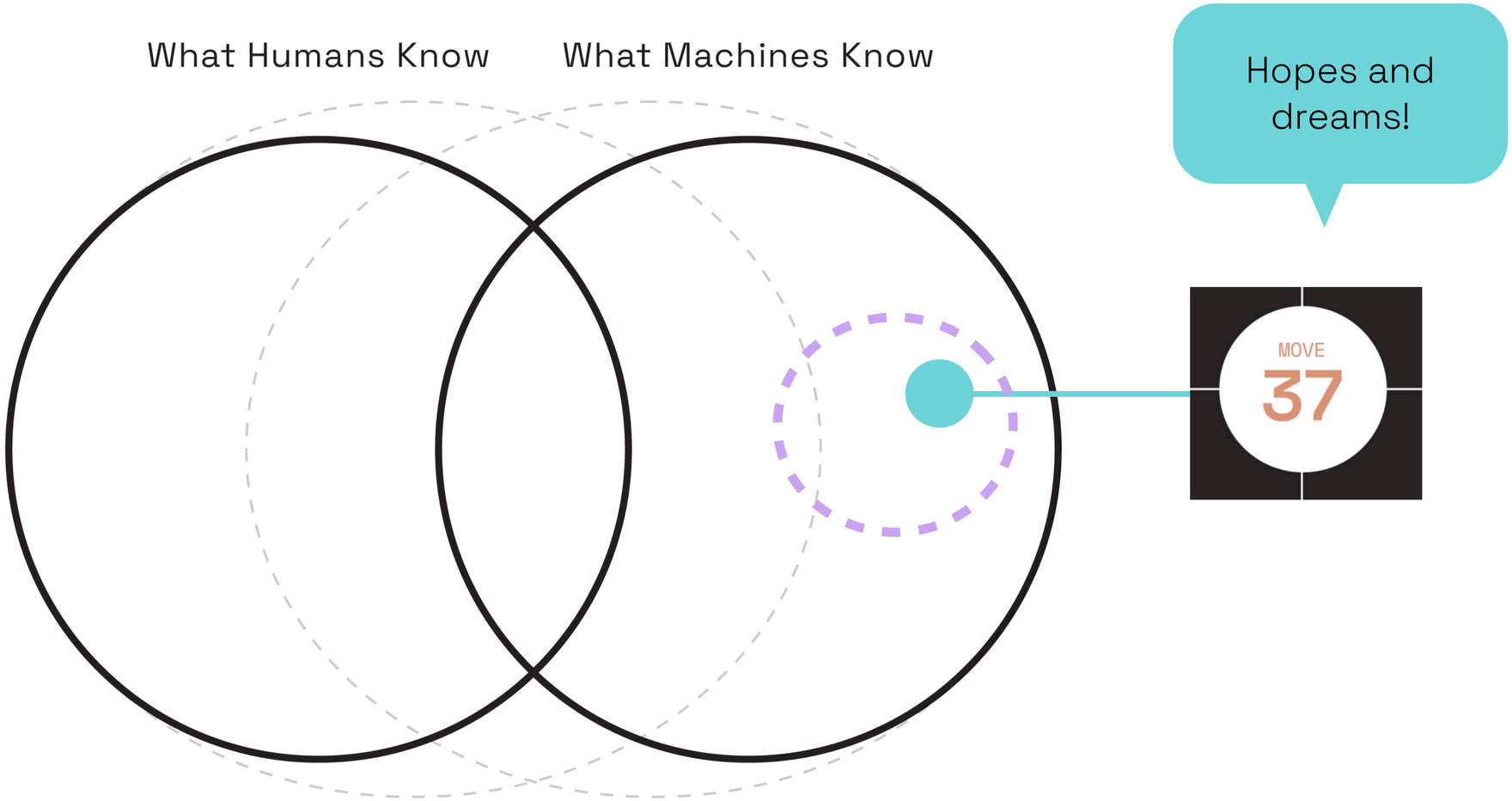
1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Studying **how humans use this tool** (what can go wrong) is as important as making the tool
[Poursabzi-sangdeh et al. 2018], [Kaur et al. 2020]

Expectation gap will be different **depending on what you are trying to do,** the end-task.
Most existing methods don't do what you think they do.
Select simplest option that directly achieves your goal.

Studying machines as if they are a new species: Observational and controlled studies.

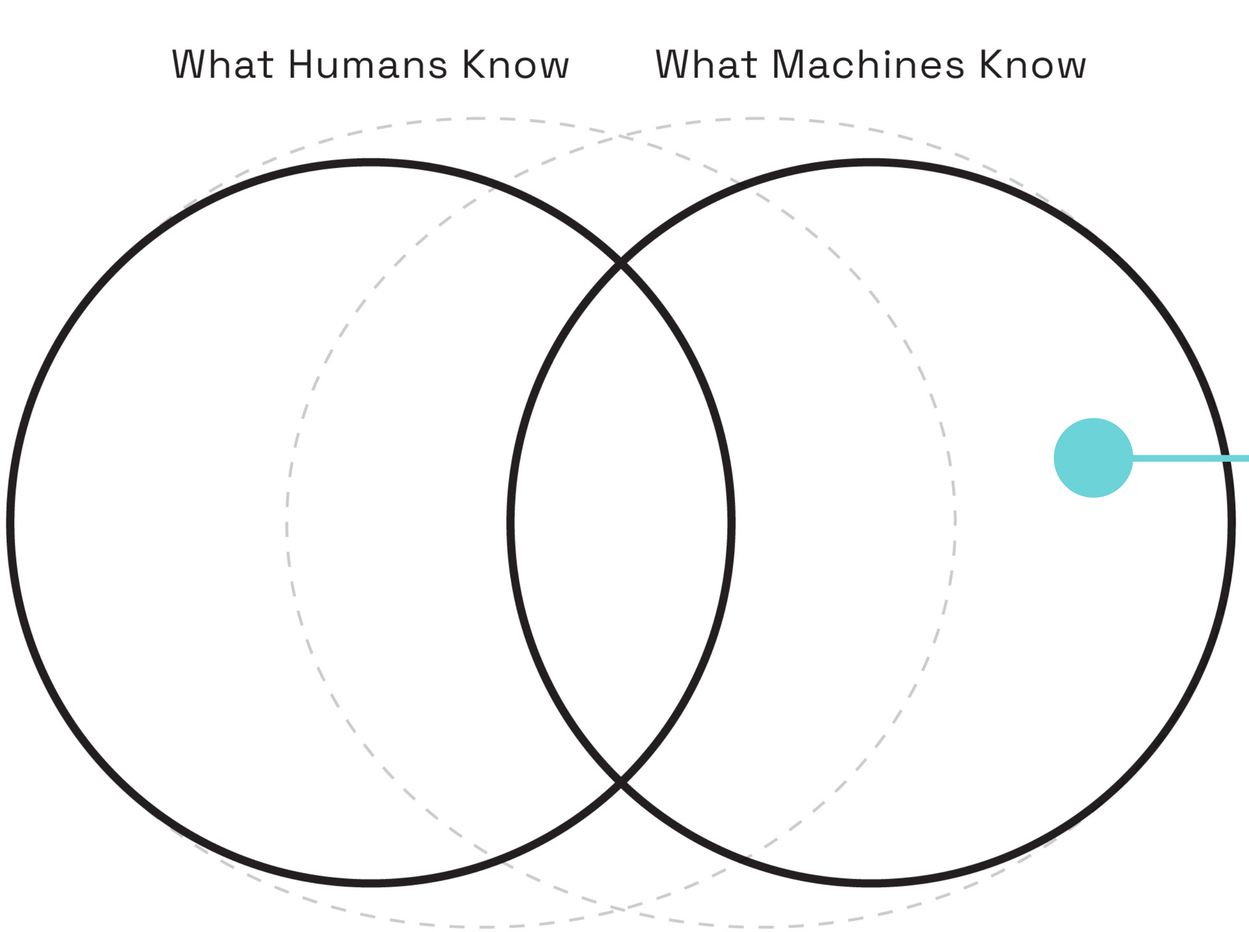
Coming back to our ultimate dream...



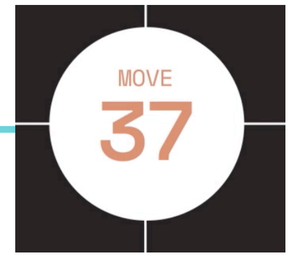
Ways to study the new species

Observational study:
Given data, discover behavior

controlled study:
Intervene, observe, discover.



Hopes and dreams!

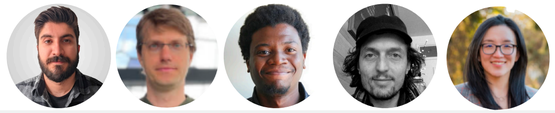


Ways to study the new species

[Neurips 2022]

Beyond Rewards: a Hierarchical Perspective on Offline Multiagent Behavioral Analysis

Shayegan Omidshafiei somidshafiei@google.com
Andrei Kapishnikov kapishnikov@google.com
Yannick Assogba yassogba@google.com
Lucas Dixon ldixon@google.com
Been Kim beenkim@google.com



[submitted, 2023]

Concept-based Understanding of Emergent Multi-Agent Behavior

Niko A. Grupen [1 *] **Natasha Jaques** [2] **Been Kim** [2] **Shayegan Omidshafiei** [2]

hat Machines Know

Observational study:
Given data, discover behavior

controlled study:
Intervene, observe, discover.

Hopes and dreams!

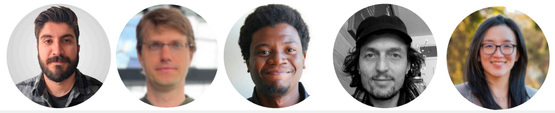
MOVE 37

Ways to study the new species

[Neurips 2022]

Beyond Rewards: a Hierarchical Perspective on Offline Multiagent Behavioral Analysis

Shayegan Omidshafiei somidshafiei@google.com
Andrei Kapishnikov kapishnikov@google.com
Yannick Assogba yassogba@google.com
Lucas Dixon ldixon@google.com
Been Kim beenkim@google.com



[submitted, 2023]

Concept-based Understanding of Emergent Multi-Agent Behavior

Niko A. Grupen [1 *] Natasha Jaques [2] Been Kim [2] Shayegan Omidshafiei [2]

hat Machines Know

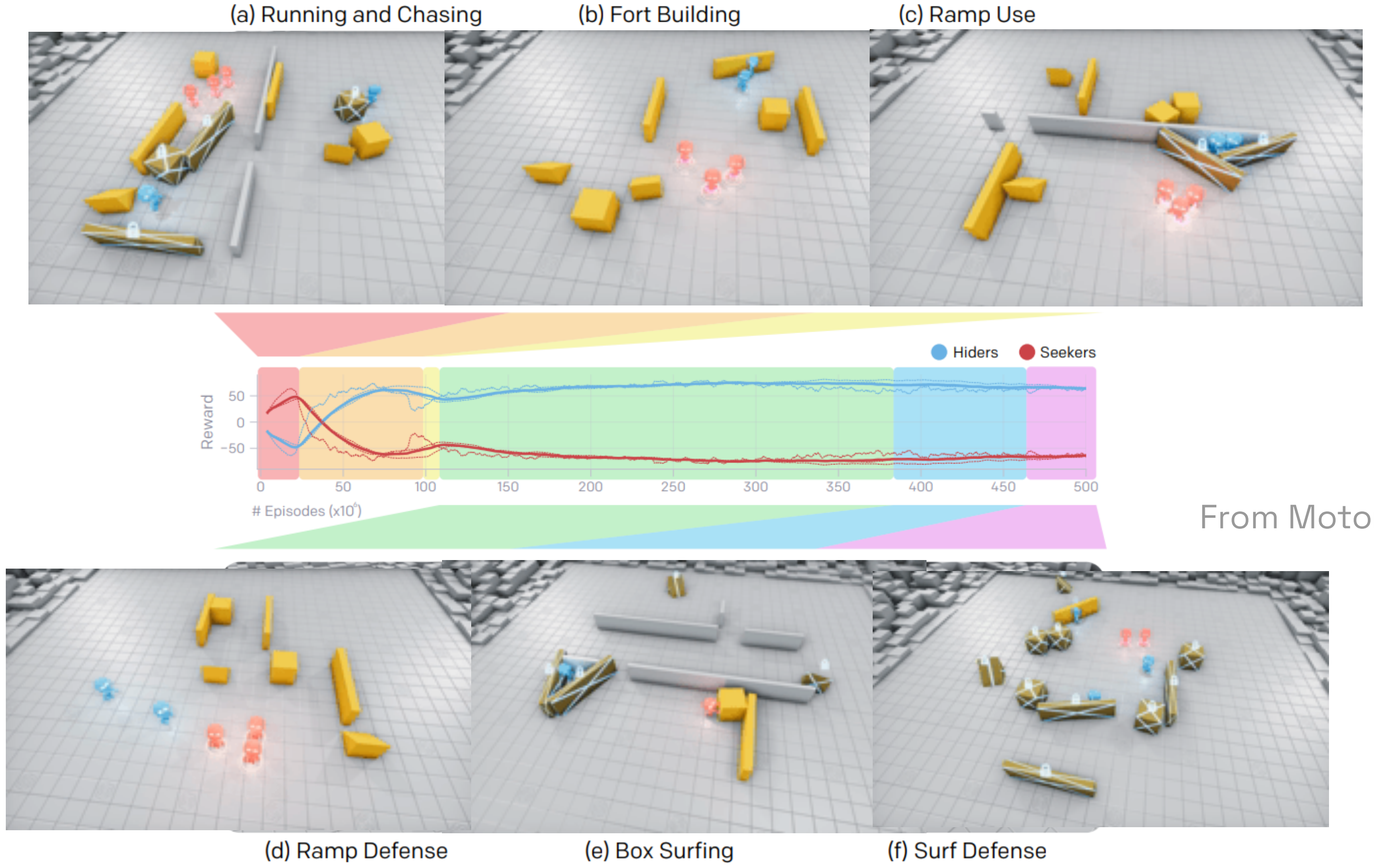
Observational study:
Given data, discover behavior

controlled study:
Intervene, observe, discover.

Hopes and dreams!

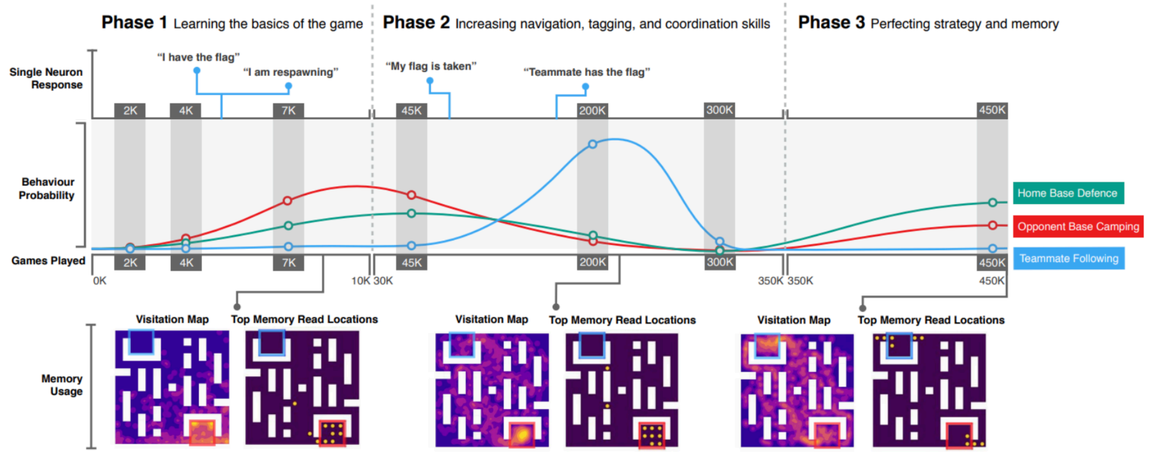
MOVE 37

Emergent behaviors in multi-agent system are fascinating - but require manual labeling



Emergent Tool Use From Multi-Agent Autocurricula (Baker et al., 2019)

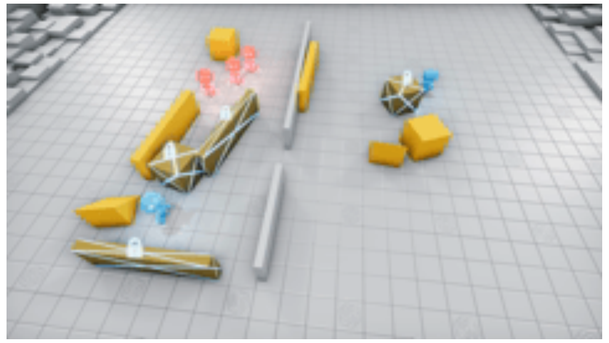
From Motor Control to Team Play in Simulated Humanoid Football (Liu et al., 2021)



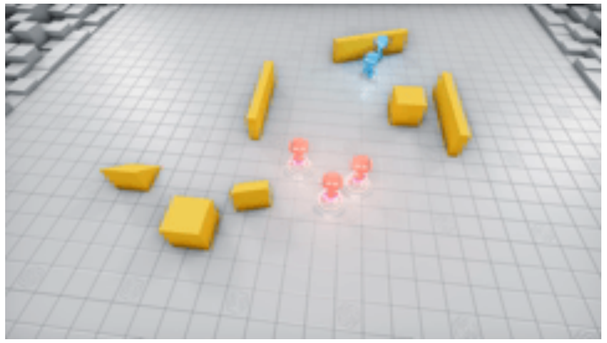
Human-level performance in 3D multiplayer games with population-based reinforcement learning (Jaderberg et al., 2019)

Emergent behaviors in multi-agent system are fascinating - but require manual labeling

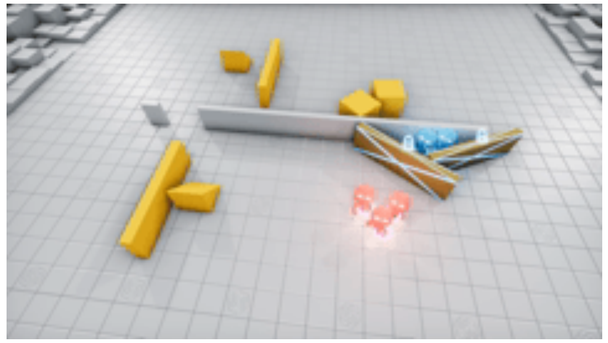
(a) Running and Chasing

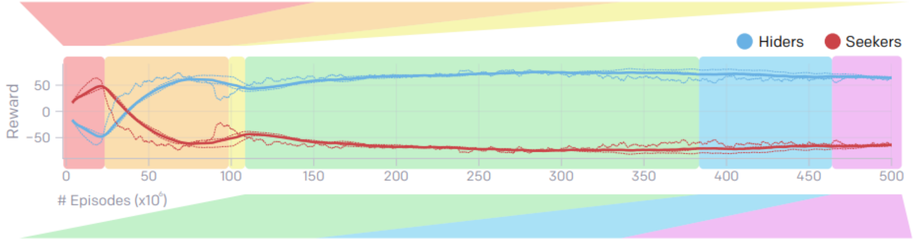


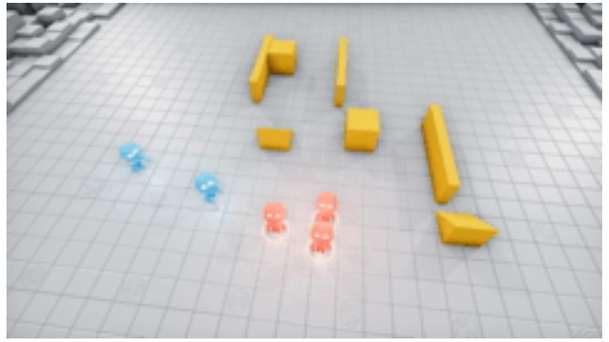
(b) Fort building



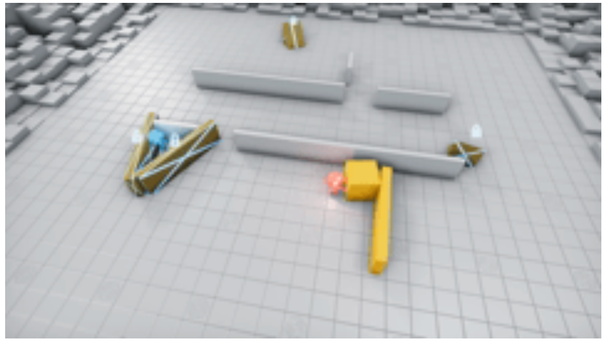
(c) Ramp Use



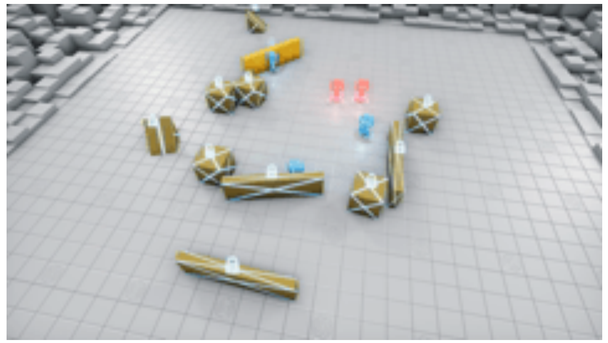




(d) Ramp Defense



(e) Box Surfing

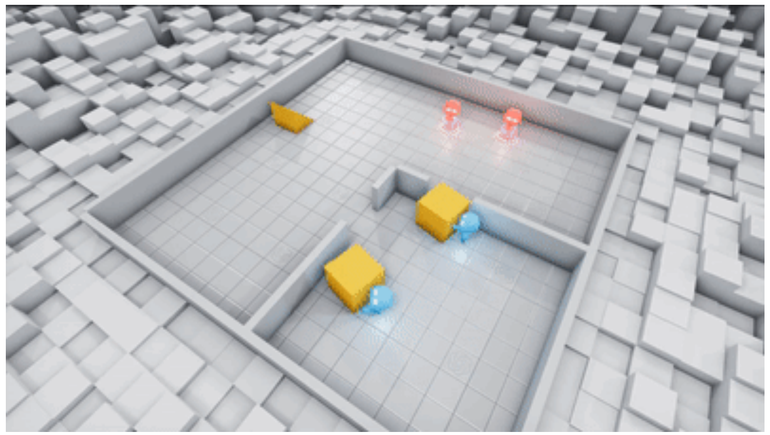


(f) Surf Defense

Emergent Tool Use From Multi-Agent Autocurricula (Baker et al., 2019)

Observational study: Given **only** observational data, discover emergent multi-agent behaviors.

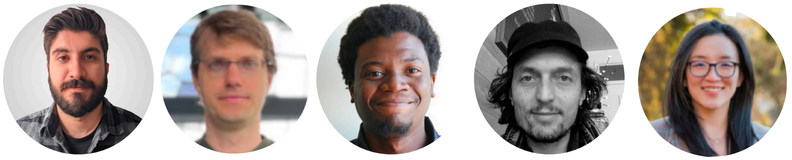
- Objective: automate the detection & visualization of multi-agent behaviors
- Problem:
  - Given state-action trajectories for N agents $\tau = \left(s_0, (a_0^i)_{i\in\mathcal{I}}, \ldots, s_{T-1}, (a_{T-1}^i)_{i\in\mathcal{I}}, s_T\right)$
  - Discover agent behavior clusters given data $\mathcal{D} = \{\tau_1, \ldots, \tau_K\}$



[Neurips 2022]

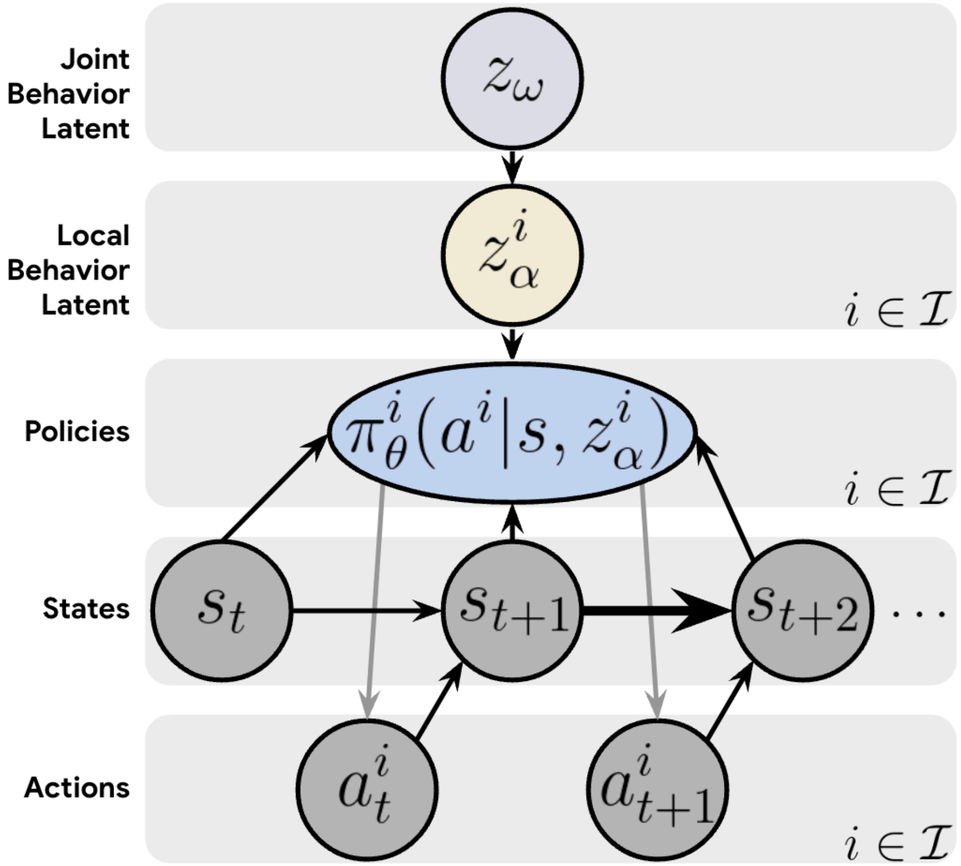
Beyond Rewards: a Hierarchical Perspective on Offline Multiagent Behavioral Analysis

Shayegan Omidshafiei somidshafiei@google.com
Andrei Kapishnikov kapishnikov@google.com
Yannick Assogba yassogba@google.com
Lucas Dixon ldixon@google.com
Been Kim beenkim@google.com



Multiagent Offline Hierarchical Behavior Analyzer (MOHBA)

- Method: Learn the joint + local latent distributions & latent-conditioned policies that maximize the above in expectation, and identify interesting behaviors in this learned space



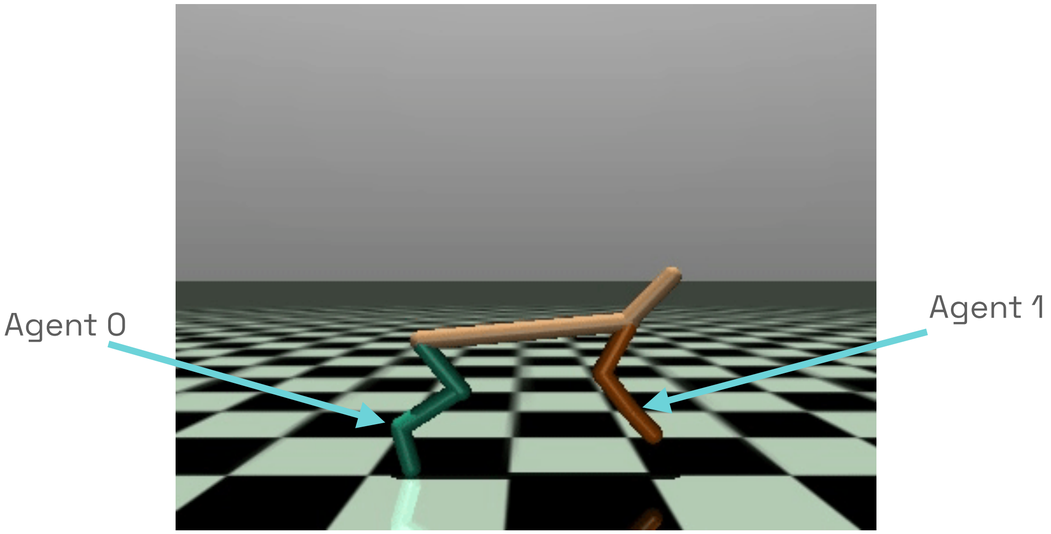
Variational lower bound

$$J_{lb} = \mathbb{E}_{\tau\sim\mathcal{D}, z_\alpha\sim q_\phi(z_\alpha|\tau)}\left[\sum_{t,i}\log\pi^i_\theta(a^i_t|s_t, z^i_\alpha)\right]$$

$$-\beta\left[\mathbb{E}_{\tau\sim\mathcal{D}, z_\omega\sim q_\phi(z_\omega|\tau)}\left[\sum_i D_{\mathrm{KL}}(q_\phi(z^i_\alpha|\tau)||p_\theta(z^i_\alpha|z_\omega))\right]\right.$$

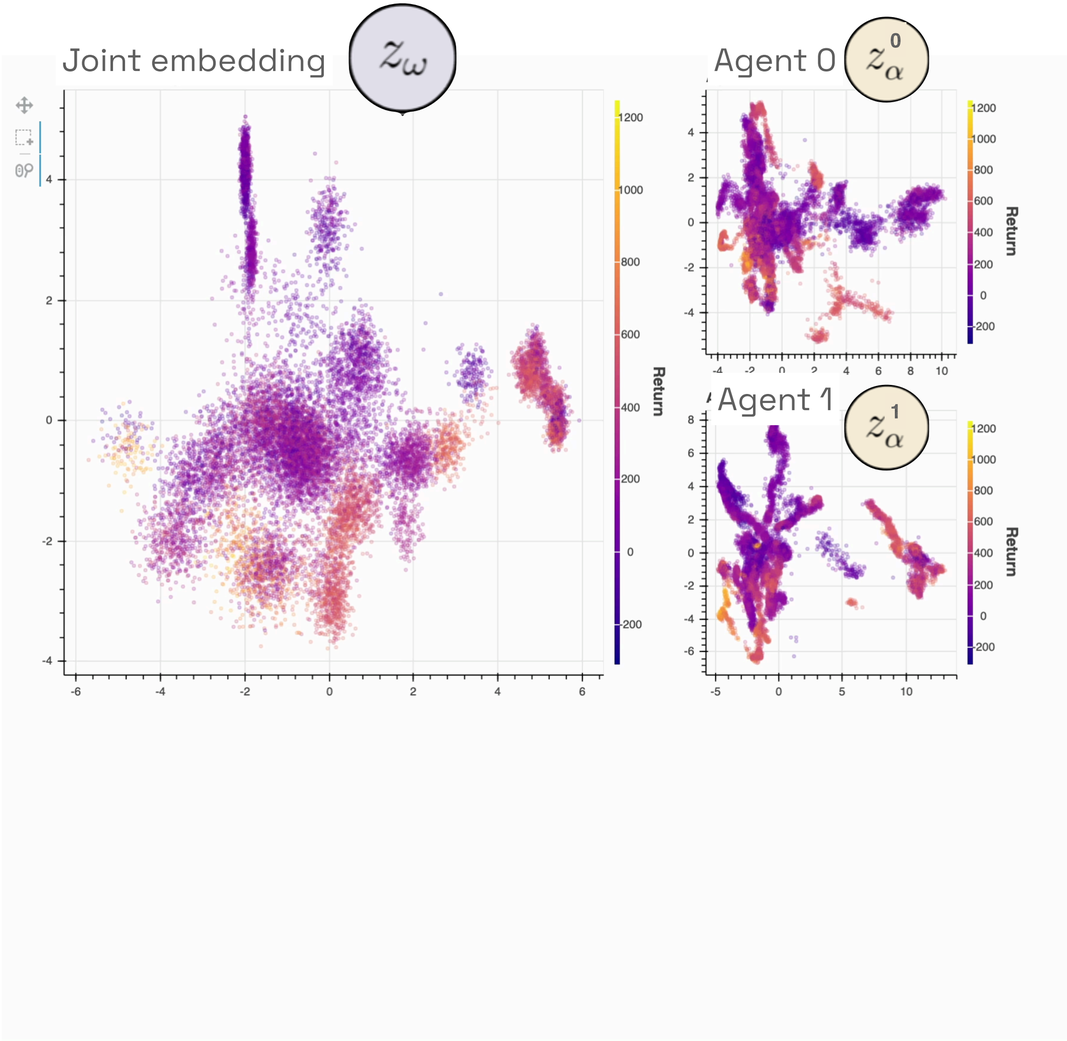
$$\left.+\mathbb{E}_{\tau\sim\mathcal{D}}\left[D_{\mathrm{KL}}(q_\phi(z_\omega|\tau)||p_\theta(z_\omega))\right]\right],$$

MOHBA on MuJoCo



Online interactive visualization

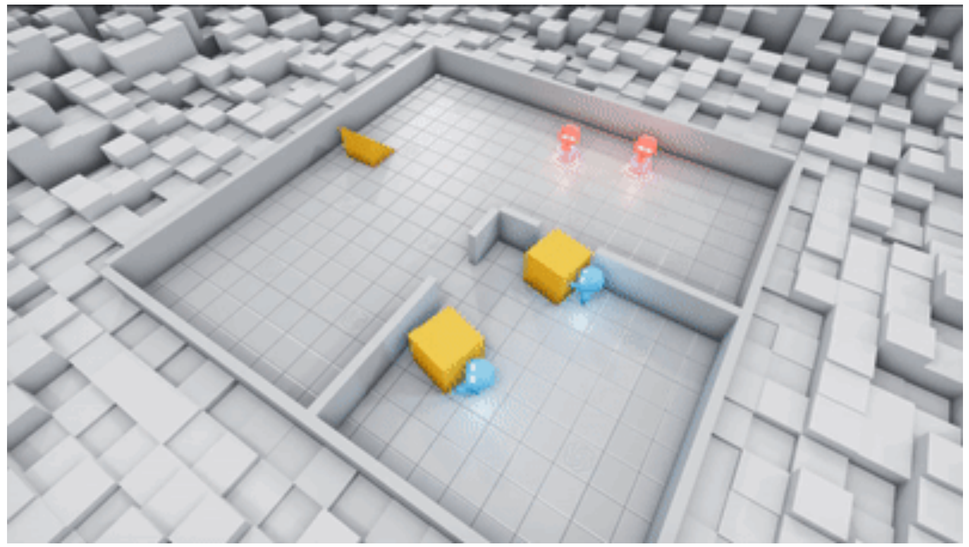
Visit here



[1]Environment codebase https://github.com/schroederdewitt/multiagent_mujoco

MOHBA on OpenAI's Hide and Seek

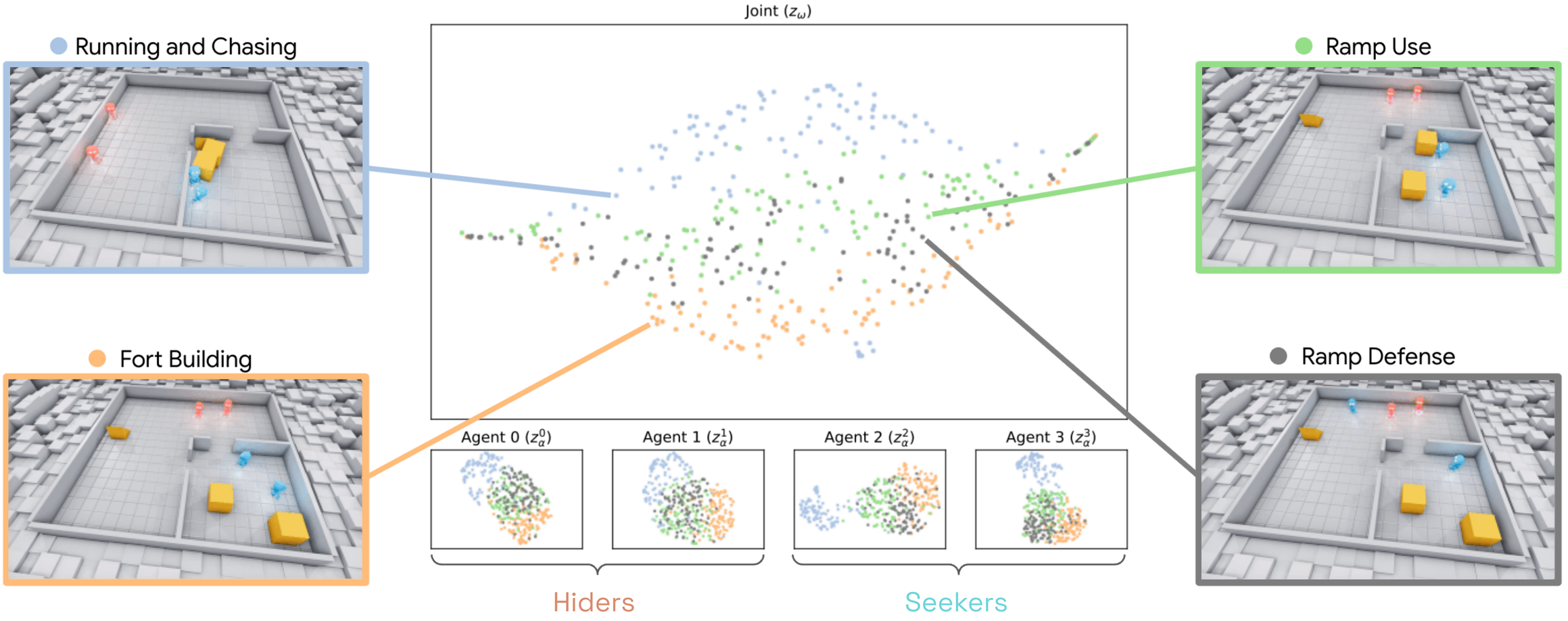
- Domain overview:
  - 4 agents (2 hiders, 2 seekers) + interactive ramps / boxes
  - Large action-observation space:
    - 100-dimensional observations
    - 5-dimensional actions
  - 200 decision-steps (15 action-repeats per step)
- MOHBA on shuffled trajectories from 4 human-annotated policies (open-sourced by OpenAI)



[1]Codebase and policies https://github.com/openai/multi-agent-emergence-environments

MOHBA on OpenAI's Hide and Seek

- MOHBA's results recover behaviors manually tagged by humans (labeled by OpenAI)



Summary and dirty laundry

Q. Can we learn some interesting emerging behaviors of multi-agent system simply by observing them?
A. Yes.

Future work/Laundry

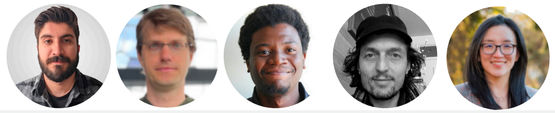
- This methods doesn't give you the names of each cluster - humans would have to explore and come up with names.
- If a cluster represents very complex behaviors that are too hard for humans to understand, this method isn't designed to assist that.
- If you have access to rewards and internal representations, you should use it! This method assumes you don't have them.

Ways to study the new species

[Neurips 2022]

Beyond Rewards: a Hierarchical Perspective on Offline Multiagent Behavioral Analysis

Shayegan Omidshafiei somidshafiei@google.com
Andrei Kapishnikov kapishnikov@google.com
Yannick Assogba yassogba@google.com
Lucas Dixon ldixon@google.com
Been Kim beenkim@google.com



[submitted, 2023]

Concept-based Understanding of Emergent Multi-Agent Behavior

Niko A. Grupen [1 *] Natasha Jaques [2] Been Kim [2] Shayegan Omidshafiei [2]

hat Machines Know

Observational study:
Given data, discover behavior

controlled study:
Intervene, observe, discover.

Hopes and dreams!

MOVE 37

Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.

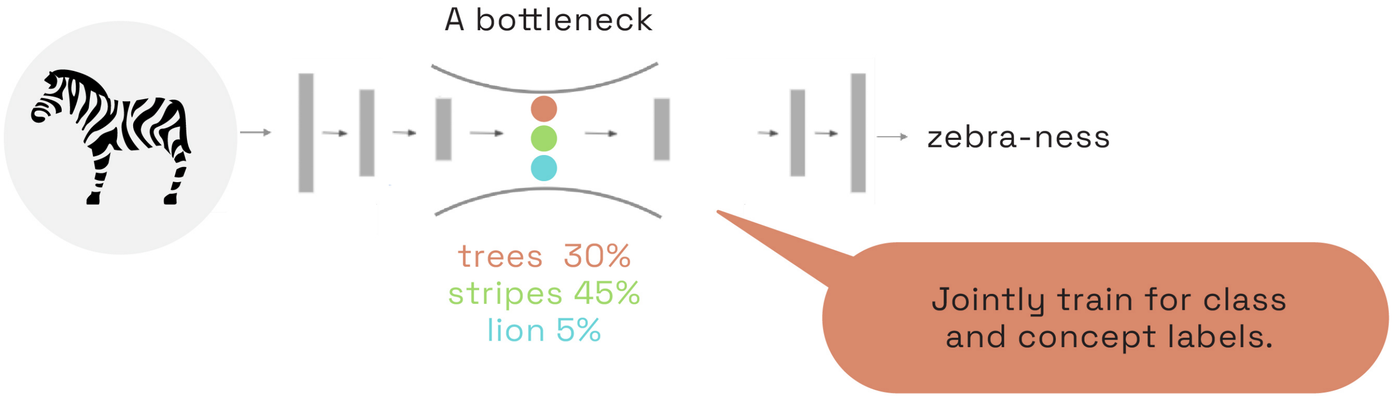
[submitted, 2023]

Concept-based Understanding of Emergent Multi-Agent Behavior

Niko A. Grupen [1 *] Natasha Jaques [2] Been Kim [2] Shayegan Omidshafiei [2]

Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.
- Approach: build in concepts as controllable units in the bottleneck.



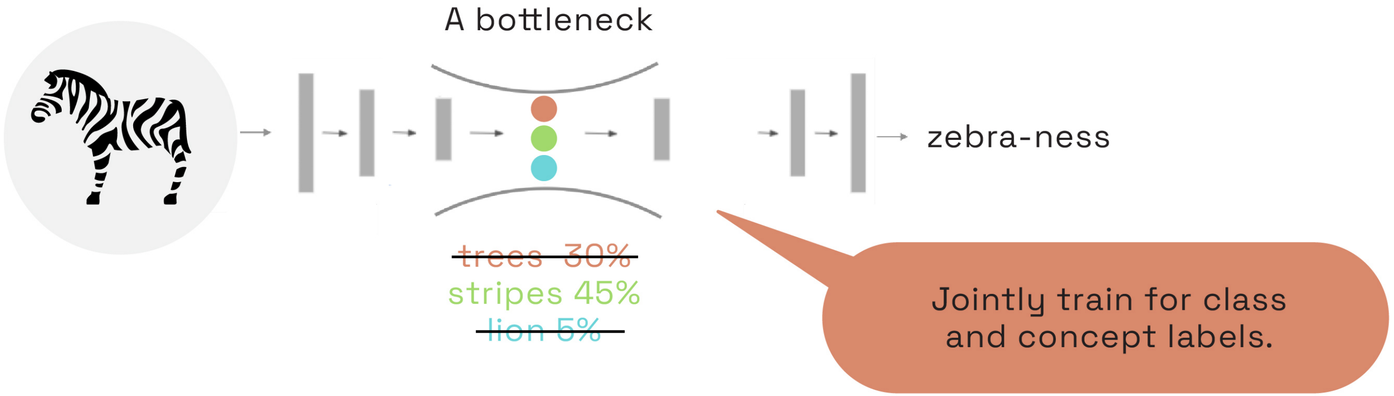
[ICML 2020]

Concept Bottleneck Models

Pang Wei Koh [*1] Thao Nguyen [*1 2] Yew Siang Tang [*1]
Stephen Mussmann [1] Emma Pierson [1] Been Kim [2] Percy Liang [1]

Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.
- Approach: build in concepts as controllable units in the bottleneck.



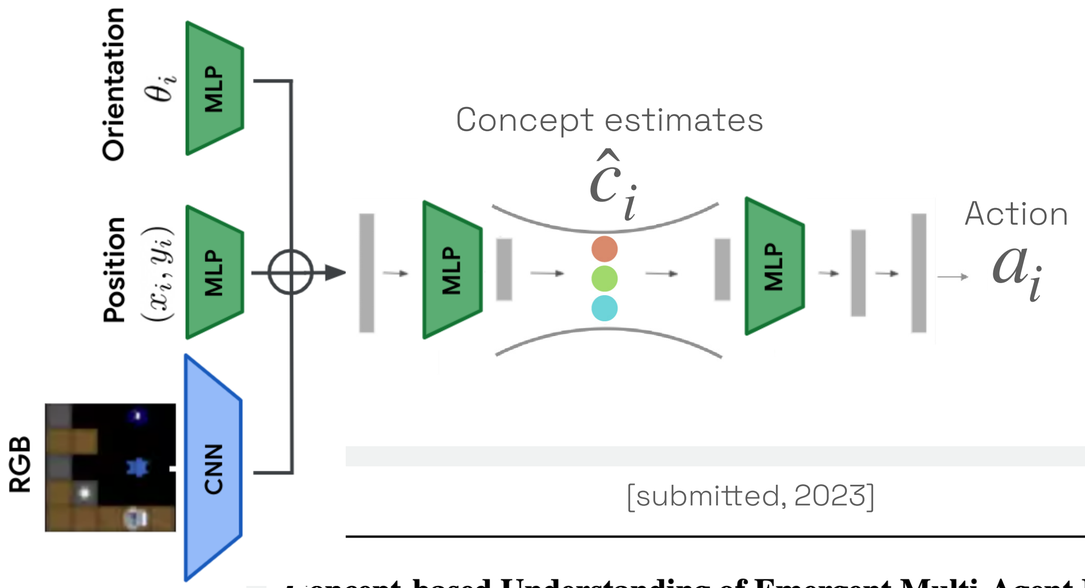
[ICML 2020]

Concept Bottleneck Models

Pang Wei Koh [*1] Thao Nguyen [*1 2] Yew Siang Tang [*1]
Stephen Mussmann [1] Emma Pierson [1] Been Kim [2] Percy Liang [1]

Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.
- Approach: build in concepts as controllable units in the bottleneck.



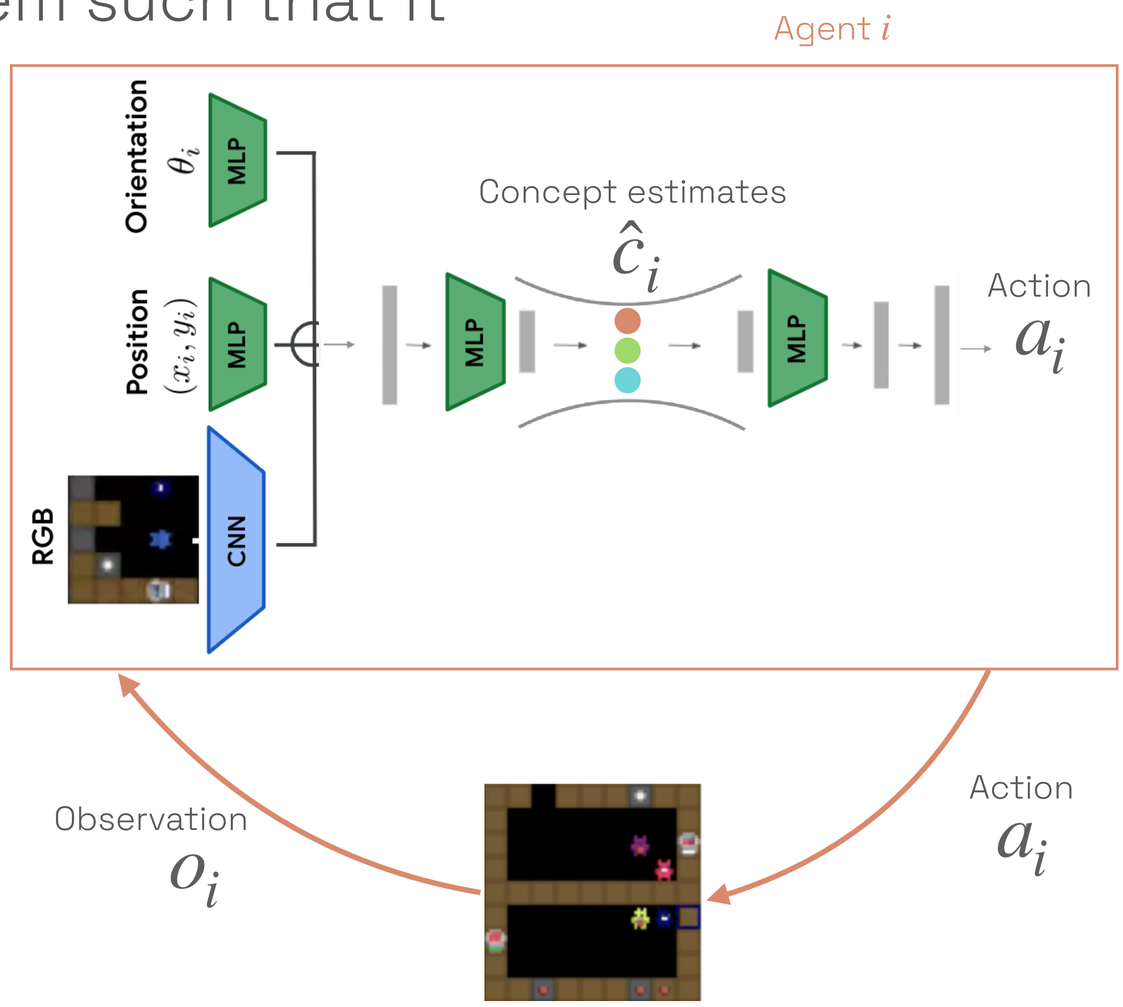
[submitted, 2023]

Concept-based Understanding of Emergent Multi-Agent Behavior

Niko A. Grupen [1 *] Natasha Jaques [2] Been Kim [2] Shayegan Omidshafiei [2]

Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.
- Approach: build in concepts as controllable units in the bottleneck.



optimize: $\mathcal{L}_{RL} + \lambda \mathcal{L}_C$

Typical PPO

Minimize c_i and $\hat{c}_i$

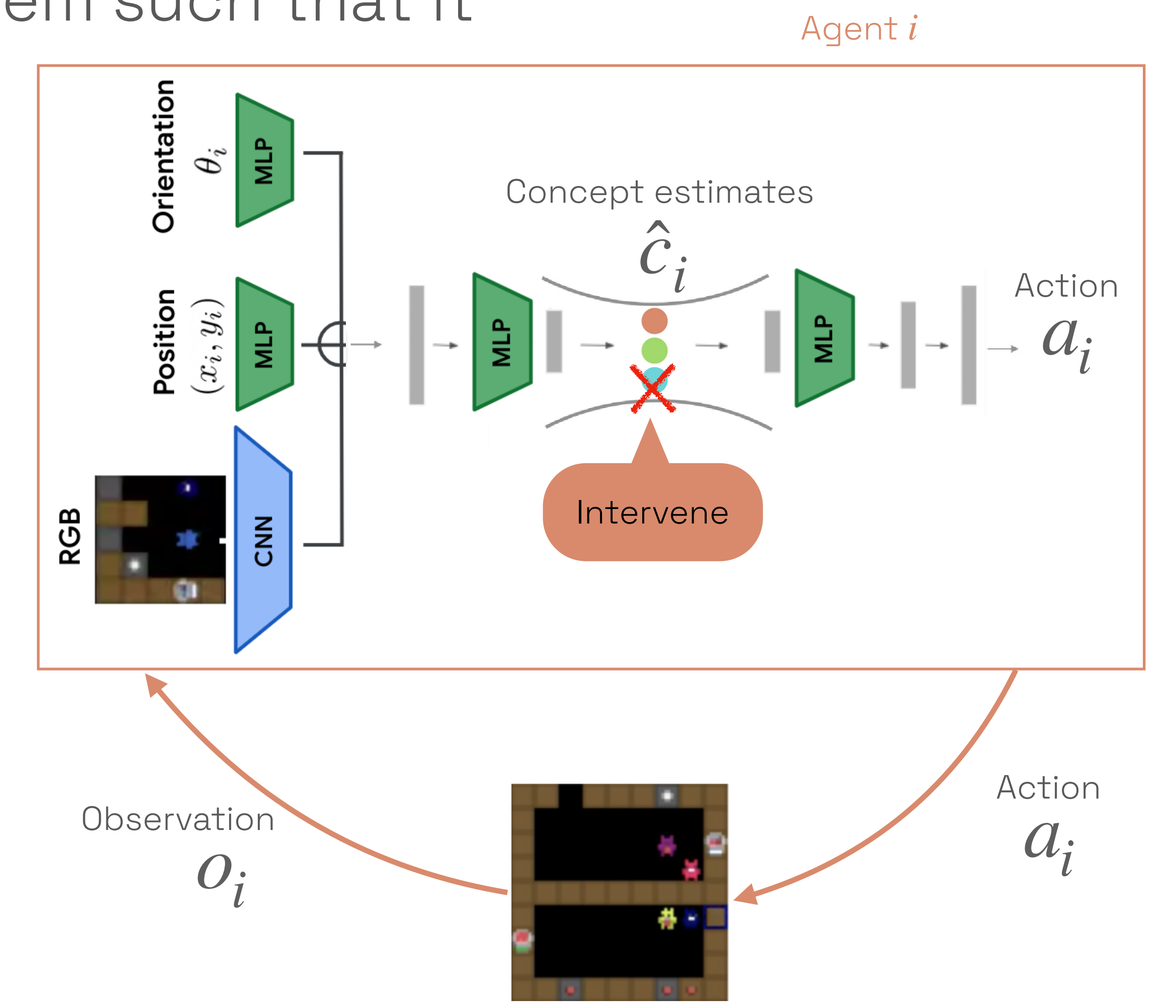
Controlled study: intervene, observe and discover emergent multi-agent behavior

- Goal: understand multi-agent behaviors via intervention.
- Problem: build a multi-agent system such that it
  - enables controlled testing
  - performs as well as baseline.
- Approach: build in concepts as controllable units in the bottleneck.

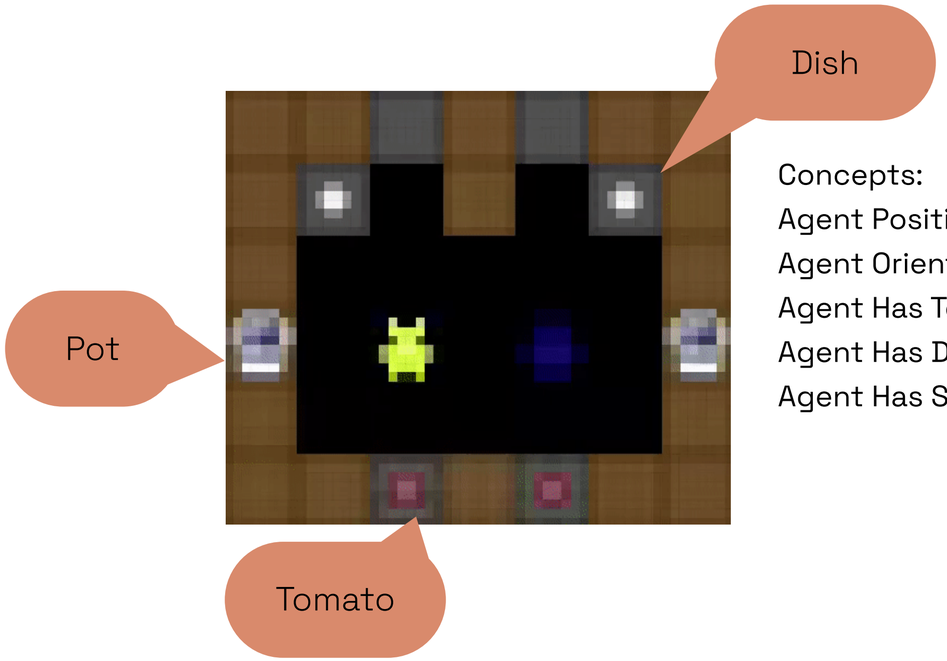
optimize: $\mathcal{L}_{RL} + \lambda \mathcal{L}_C$

Typical PPO

Minimize c_i and $\hat{c}_i$



Study1: Emerging coordination in Domain Cooking game



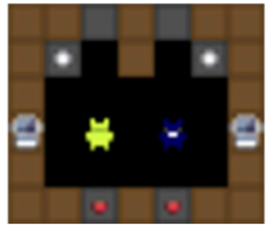
Concepts:
Agent Position
Agent Orientation
Agent Has Tomato
Agent Has Dish
Agent Has Soup

Recipe: Tomato Soup

1. Bring 3 tomatoes to a cooking pot. → Reward!
2. Wait for tomato soup to cook.
3. Bring a dish to the cooking pot and pour soup into it. → Reward!
4. Deliver soup to the delivery location.

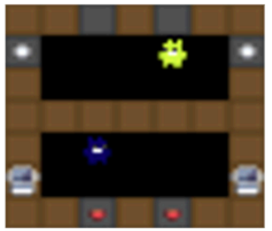
Reward!

Basic Environment



Passable Environment

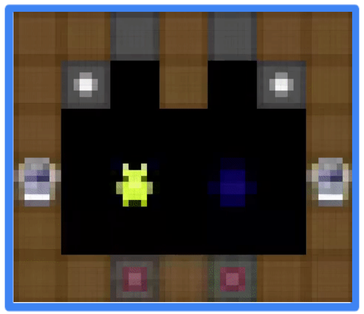
Impassable Environment

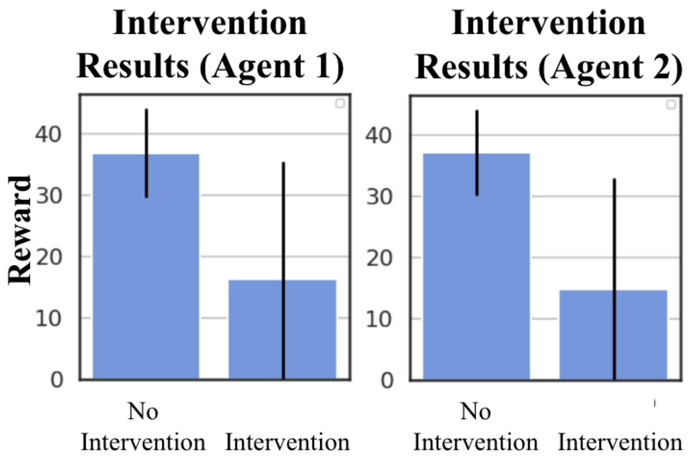


Asymmetric Environment

Study1: Emerging coordination - How much do they coordinate?

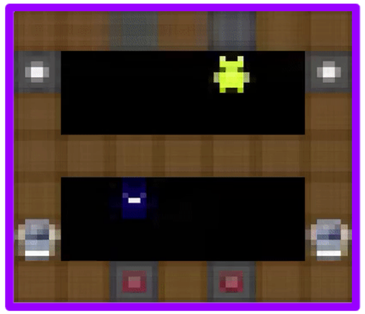
Basic Environment

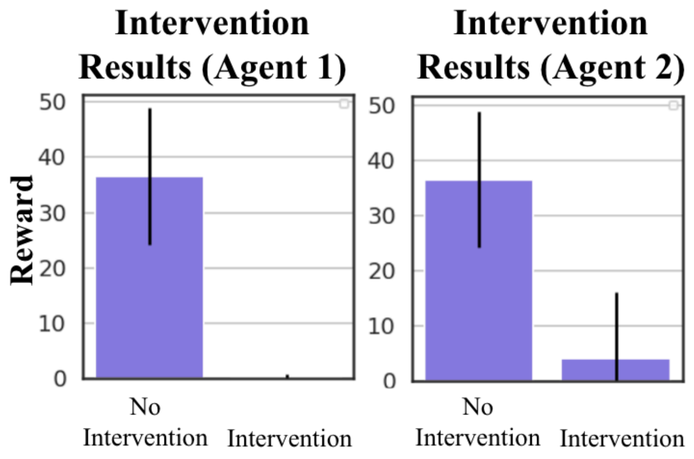




Coordination, but not forced

Impassable Environment

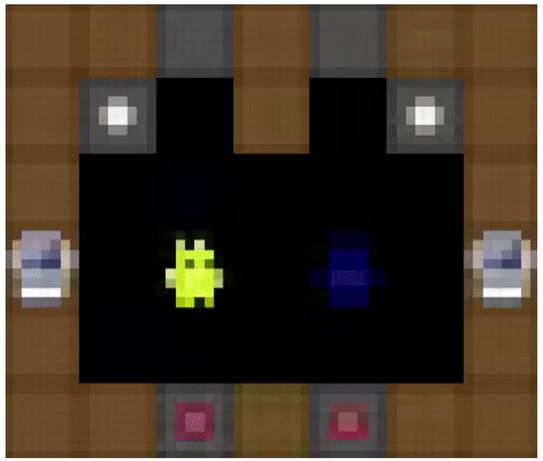


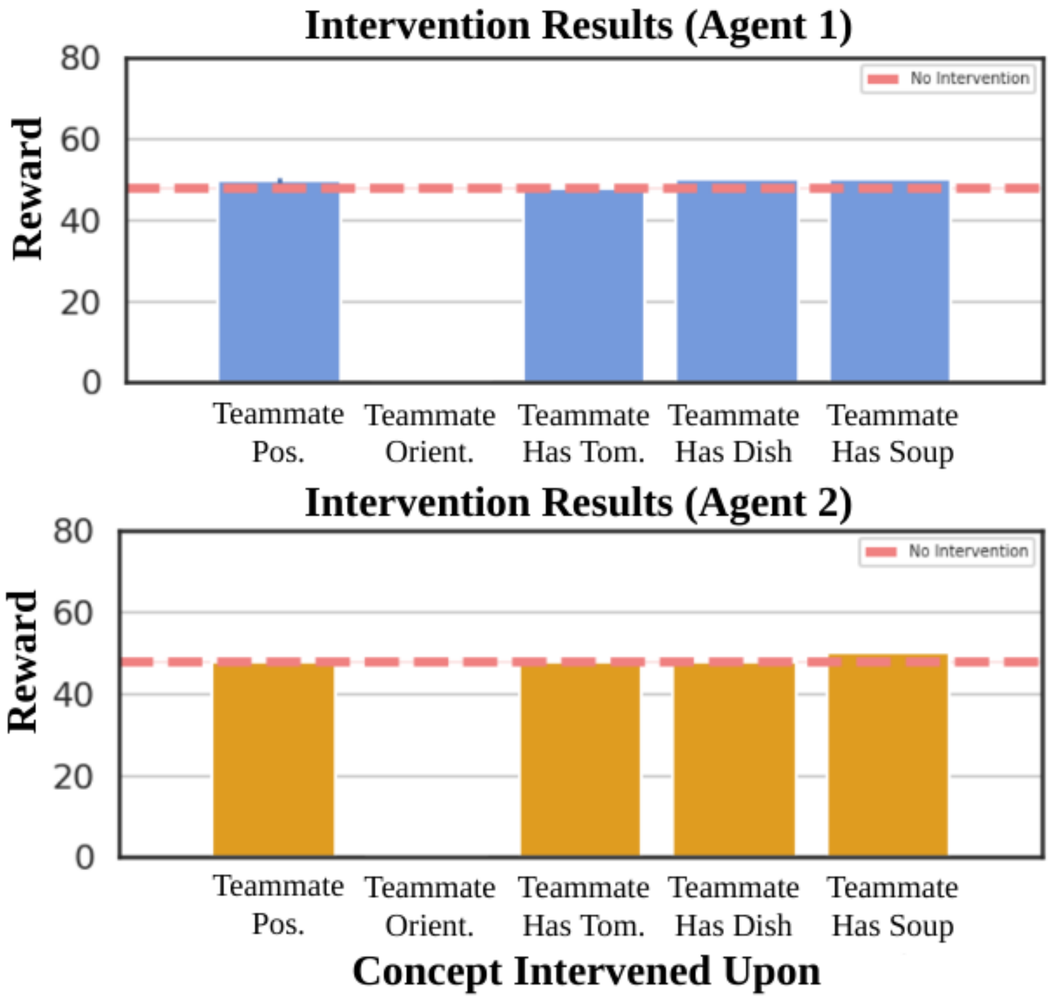


Forced coordination

Study1: Emerging coordination - What do they **use** to coordinate?

Coordination example



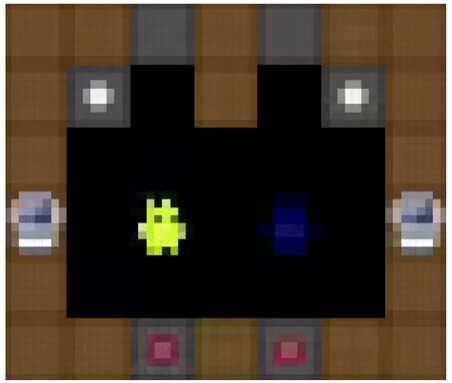


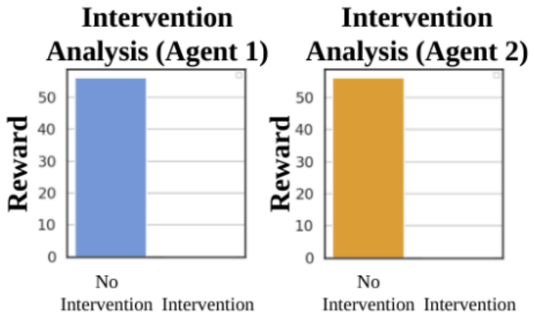
Teammate's **orientation** is the one and only factor for coordination!

Why?

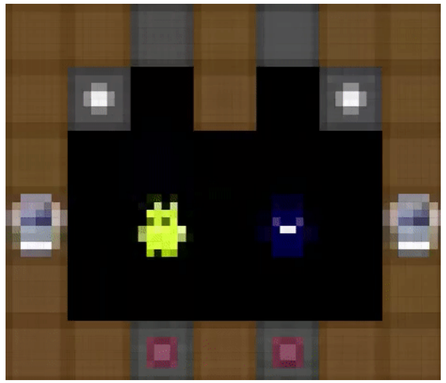
Study1: Emerging coordination - Are there free riders? - of course.

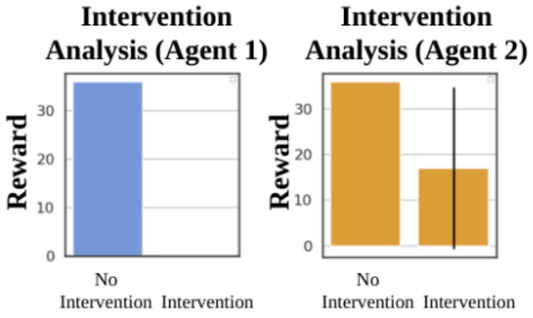
Strong Coordination



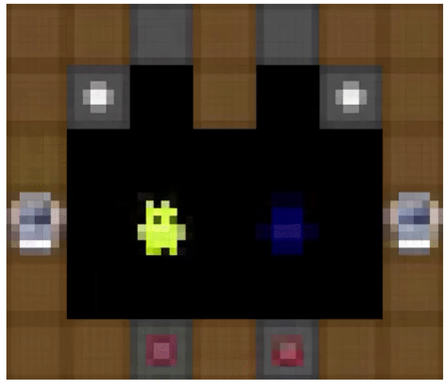


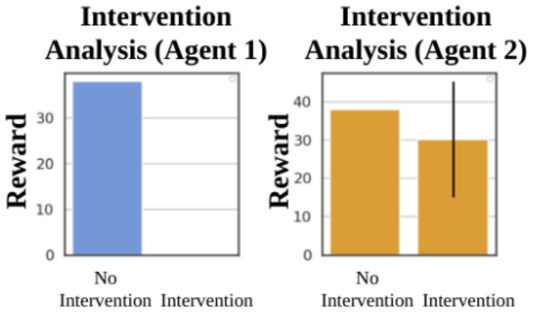
Moderate Coordination



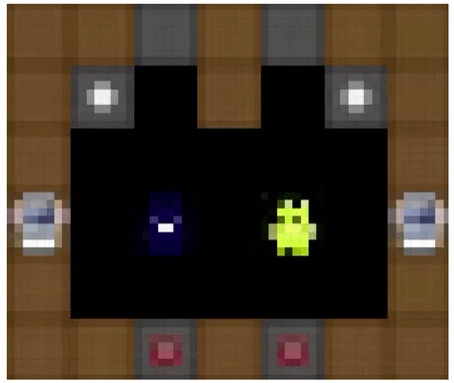


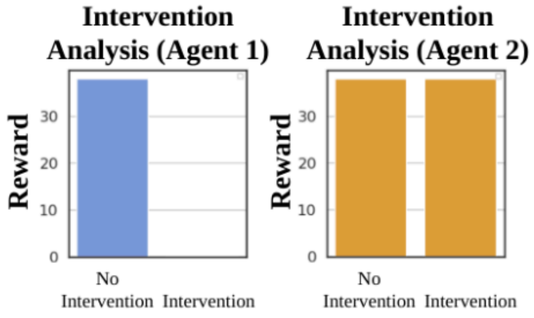
No Coordination





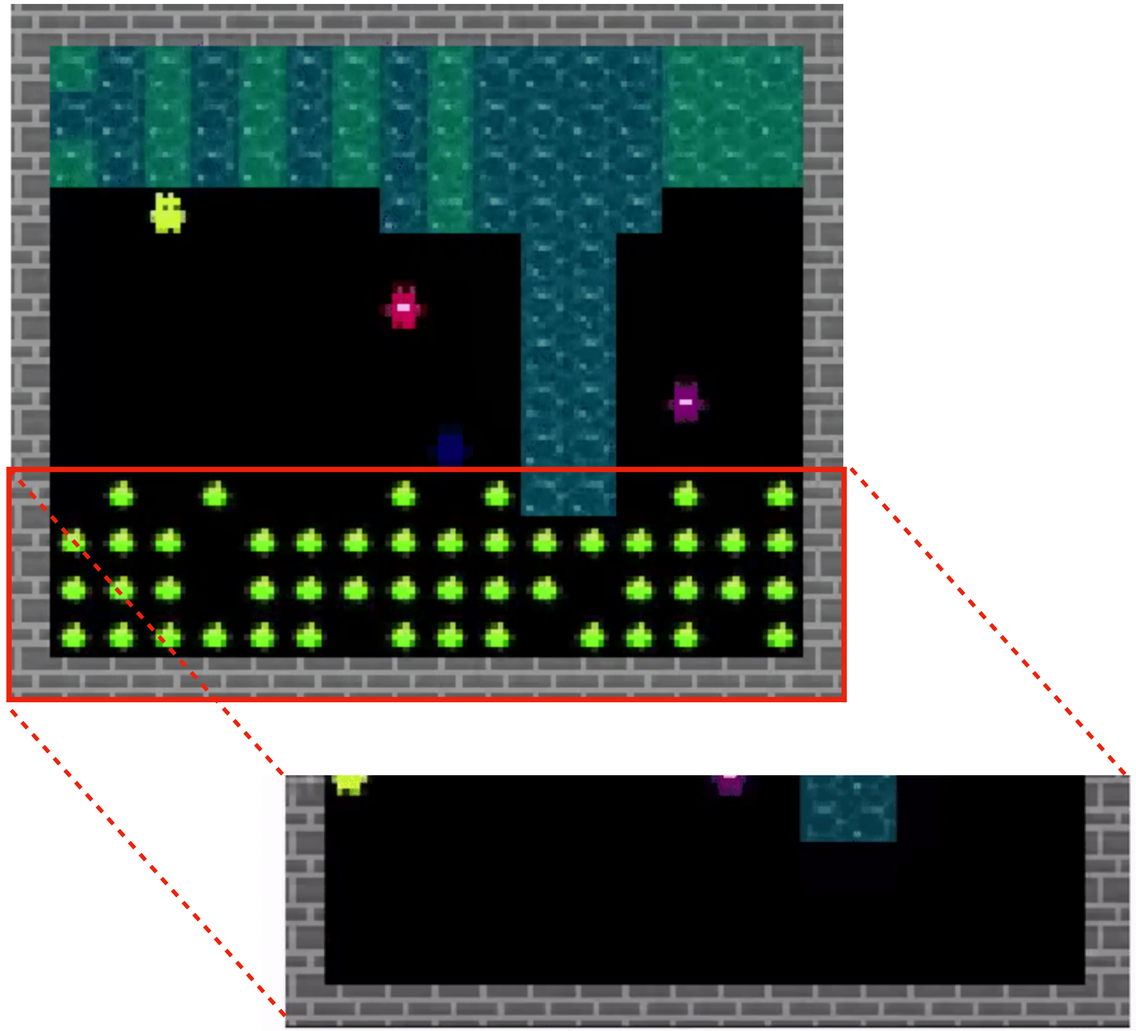
No Movement





Degree of laziness

Study2: Inter-agent social dynamics in Domain Clean Up

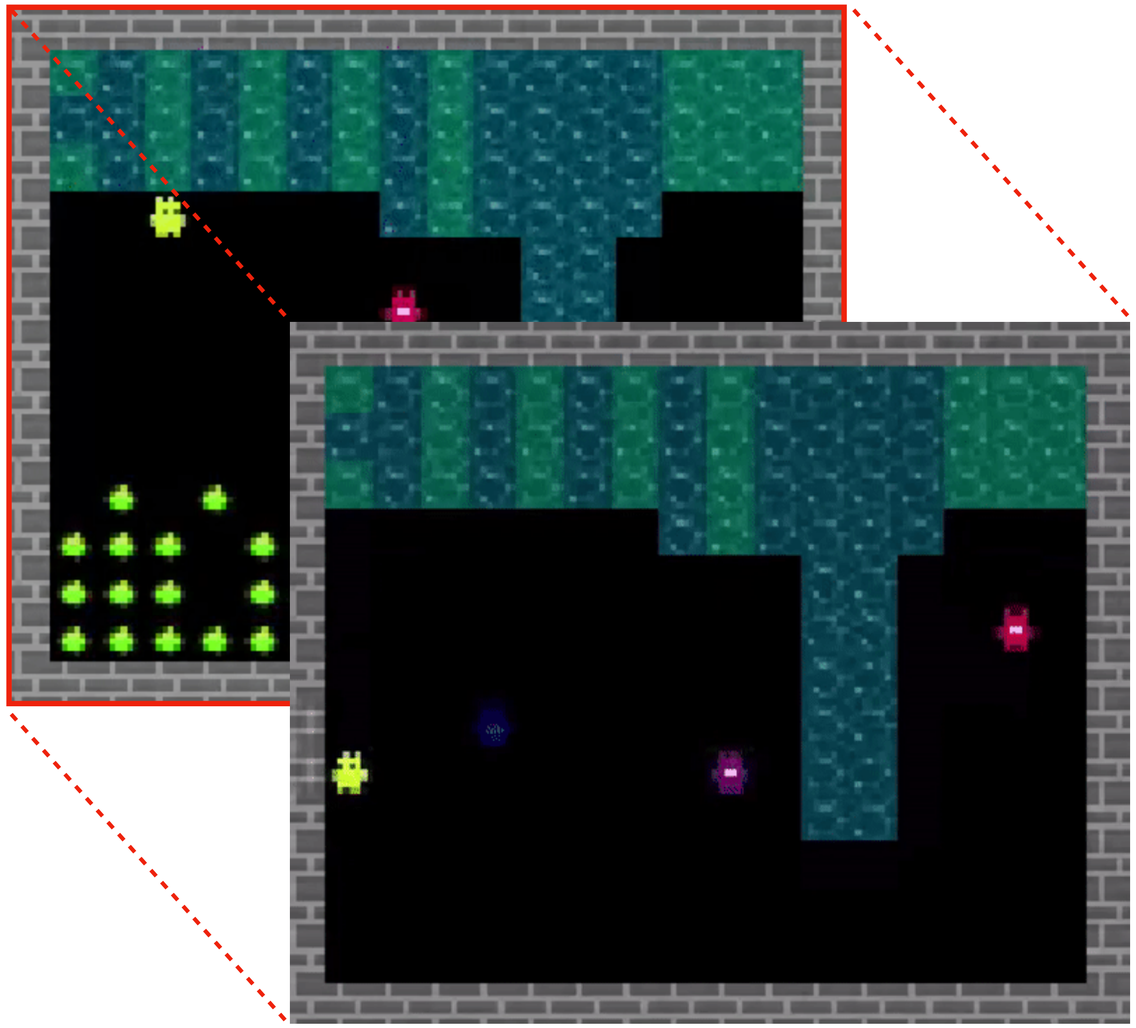


Tasks: Clean up

1 Eating apples. → **Reward!**

2 Cleaning the river.

Study2: Inter-agent social dynamics in Domain Clean Up



Tasks: Clean up

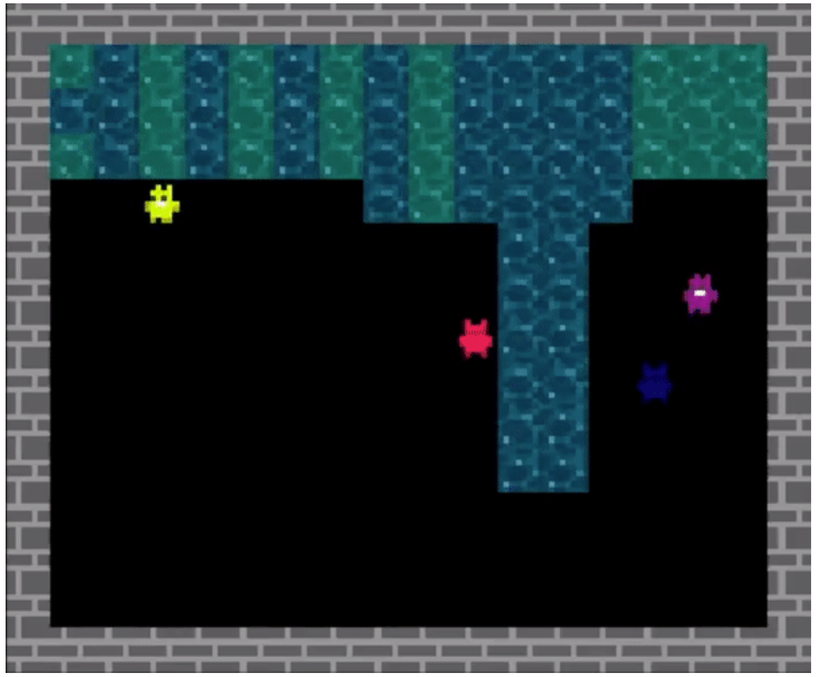
1. Eating apples. → **Reward!**
2. Cleaning the river.

Concepts:

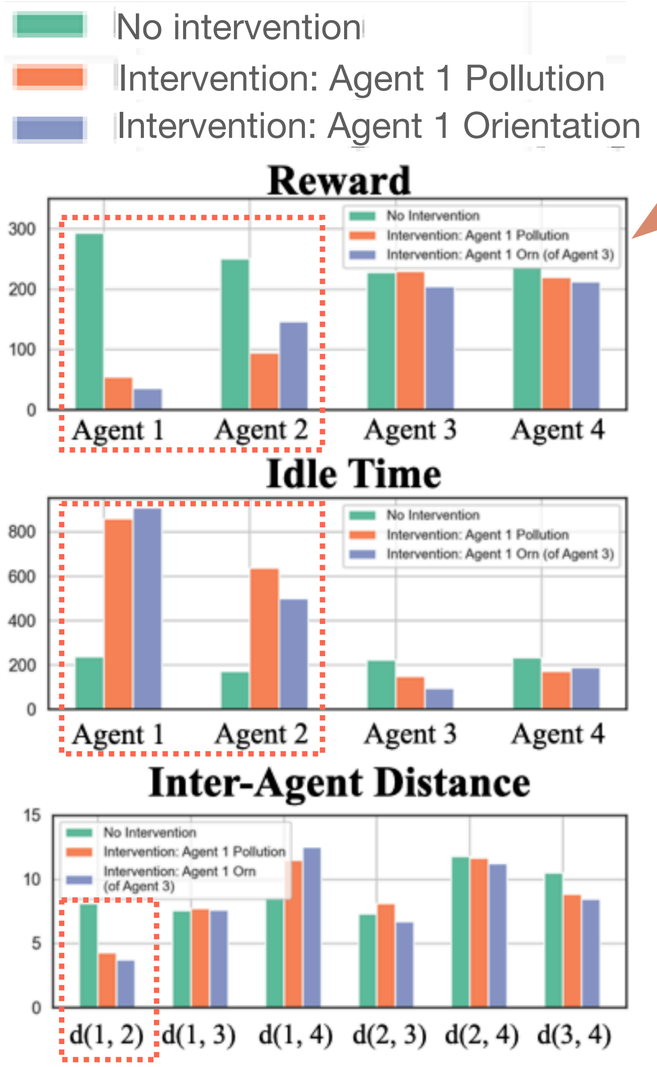
- Agent position
- Agent orientation
- Closest apple positions
- Closest pollution positions

Study2: Inter-agent social dynamics - Raw statistics alone

- Suggested story from the raw statistics: Agent 1 and 2 seem to be dependent on each other… Is this true?



Agent 1 = **Blue**
Agent 2 = **Yellow**
Agent 3 = **Pink**
Agent 4 = **Purple**



Raw stats!

Study2: Inter-agent social dynamics - building a graph using interventions

- Suggested story from the raw statistics: Agent 1 and 2 seem to be dependent on each other… Is this true?
- Let's do a bit more work: building a graph of inter-agent relationships (simplest way):

An example:

$$\boldsymbol{X} = \begin{bmatrix} x^1_{\hat{c}_j} & x^2_{\hat{c}_j} & \dots & x^n_{\hat{c}_j} \\ x^1_{\hat{c}_k} & x^2_{\hat{c}_k} & \dots & x^n_{\hat{c}_k} \\ & \vdots & & \\ x^1_{\hat{c}_l} & x^2_{\hat{c}_l} & \dots & x^n_{\hat{c}_l} \end{bmatrix}$$

Rows: movies

Columns: features of each movie (e.g., genre, length)

I-th row of X

Remaning n-1 outcomes

$$\min_{\beta_i} ||\boldsymbol{X}_i - \boldsymbol{X}_{\setminus i}\beta_i||_2^2 + \alpha||\beta_i||_1$$

Edges in the graph <-> Relationships between movies

Study2: Inter-agent social dynamics - building a graph using interventions

- Suggested story from the raw statistics: Agent 1 and 2 seem to be dependent on each other… Is this true?
- Let's do a bit more work: building a graph of inter-agent relationships (simplest way):

$$\boldsymbol{X} = \begin{bmatrix} \text{Int}(\hat{c}_j, 1) \\ \text{Int}(\hat{c}_k, 2) \\ \vdots \\ \text{Int}(\hat{c}_l, N) \end{bmatrix} = \begin{bmatrix} x^1_{\hat{c}_j} & x^2_{\hat{c}_j} & \dots & x^n_{\hat{c}_j} \\ x^1_{\hat{c}_k} & x^2_{\hat{c}_k} & \dots & x^n_{\hat{c}_k} \\ & \vdots & & \\ x^1_{\hat{c}_l} & x^2_{\hat{c}_l} & \dots & x^n_{\hat{c}_l} \end{bmatrix}$$

Intervention on concept c agent N

Intervention outcome (e.g., reward, resources collected, proximity to other agents)

I-th row of X

Remaning n-1 outcomes

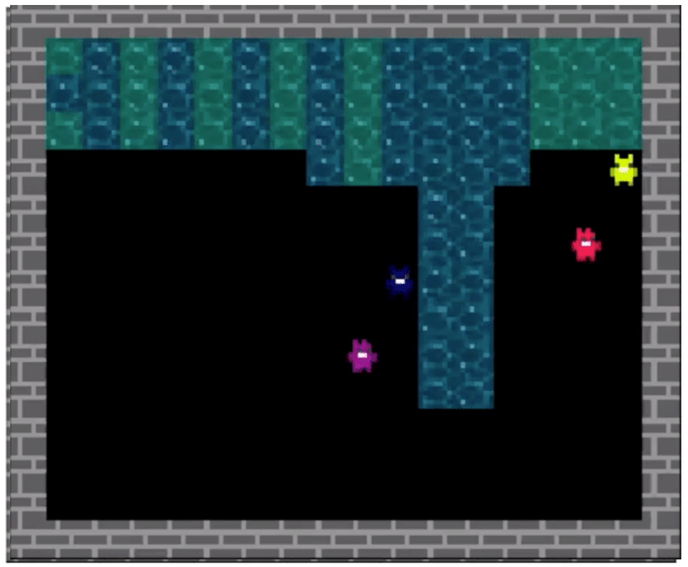
$$\min_{\beta_i} ||\boldsymbol{X}_i - \boldsymbol{X}_{\setminus i}\beta_i||_2^2 + \alpha||\beta_i||_1$$

Edges in the graph
<->
Relationships between interventions

Study2: Inter-agent social dynamics - building a graph using interventions

- Suggested story from the raw statistics: Agent 1 and 2 seem to be dependent on each other… Is this true?
- Truth: chains of social interactions
  - Agent 1's orientation is important for Agent 4. When it fails, it makes Agent 1 and 2 physically collide in the environment (accidental).
  - There is no coordination between Agent 1 and 2.

Intervention: Agent 1 Orn. (Ag3)



Agent 1 = **Blue**
Agent 2 = **Yellow**
Agent 3 = **Pink**
Agent 4 = **Purple**

Summary and dirty laundry

Q. Can we build a multi-agent system that enables interventional studies AND performs as well as baselines?

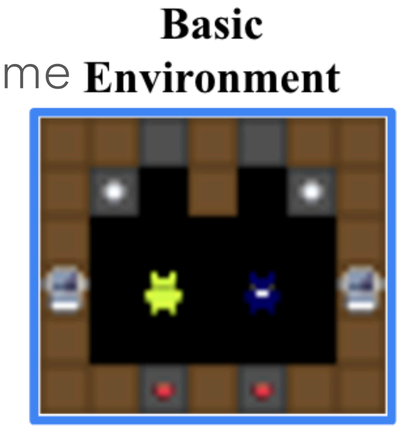
A. Yes

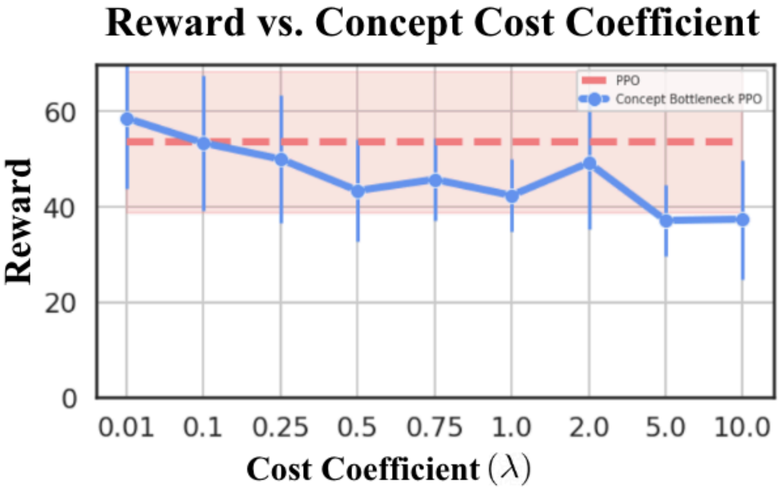
- Matches PPO performance, but it does require some tuning of λ

optimize: $\mathscr{L}_{RL} + \lambda \mathscr{L}_{C}$

Typical PPO

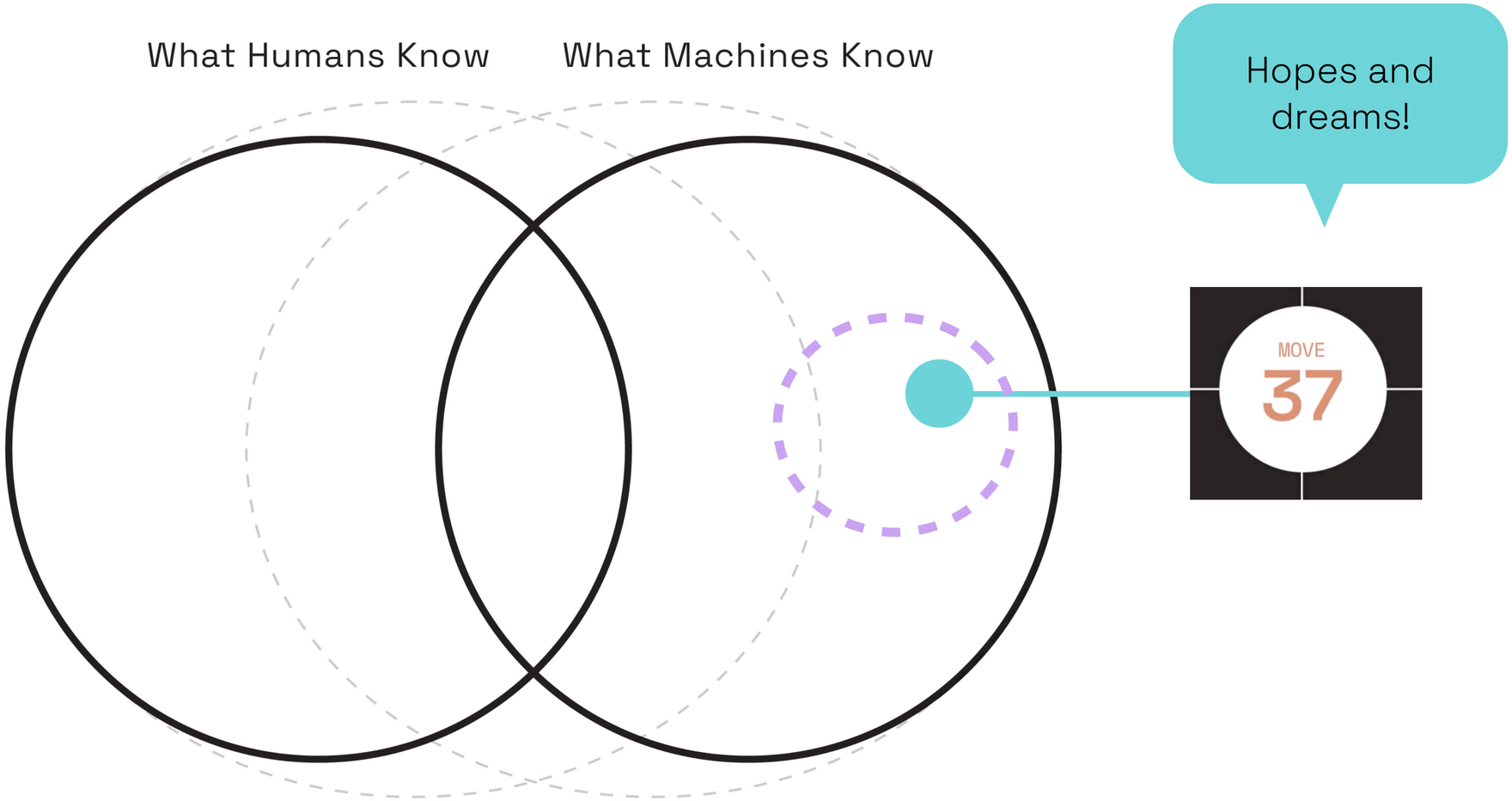
Minimize c_i and $\hat{c}_i$





- Concepts are assumed to be pre-defined and available

Looking ahead - future doesn’t need to wait.



[previous work] Studying a superhuman network: revealing alignment

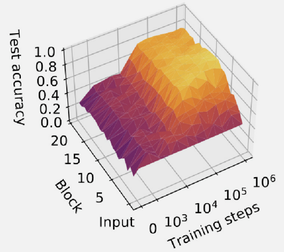
[PNAS 2022]

Acquisition of Chess Knowledge in AlphaZero

Thomas McGrath[a,1,2], Andrei Kapishnikov[b,1], Nenad Tomašev[a], Adam Pearce[b], Martin Wattenberg[b], Demis Hassabis[a], Been Kim[b], Ulrich Paquet[a], and Vladimir Kramnik[c]

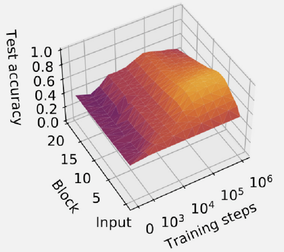
What-when-where plots: What concepts AZ learns, when and where

Can capture queen opponent



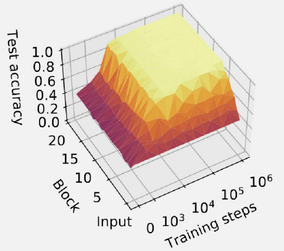
(e) Can the playing side capture their opponent's queen?

Material imbalance



(h) Past 10^5 training steps, StockFish 8's material imbalance score becomes *less* predictable from AlphaZero's later layers.

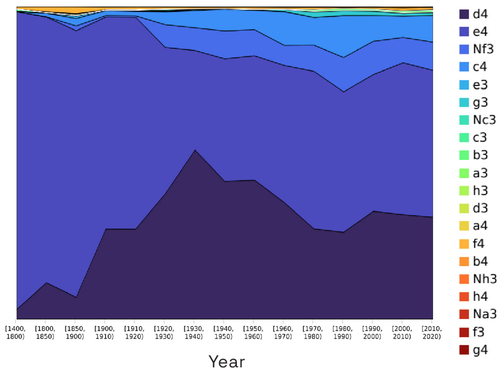
In check



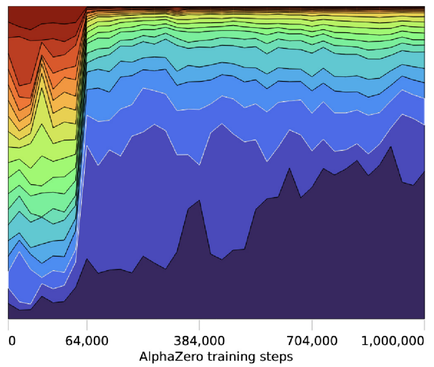
(c) Is the playing side in check?

The evolution of opening moves: Humans vs AZ

Human's (from Chessbase) opening

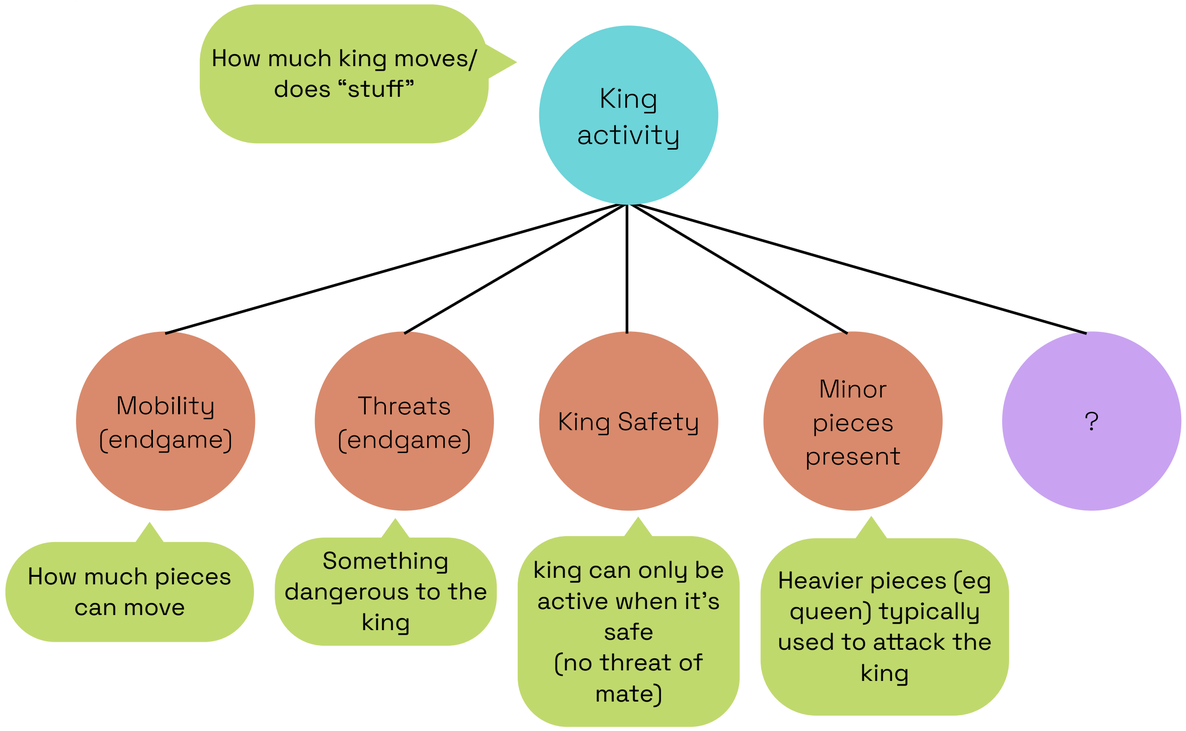


AlphaZero's opening

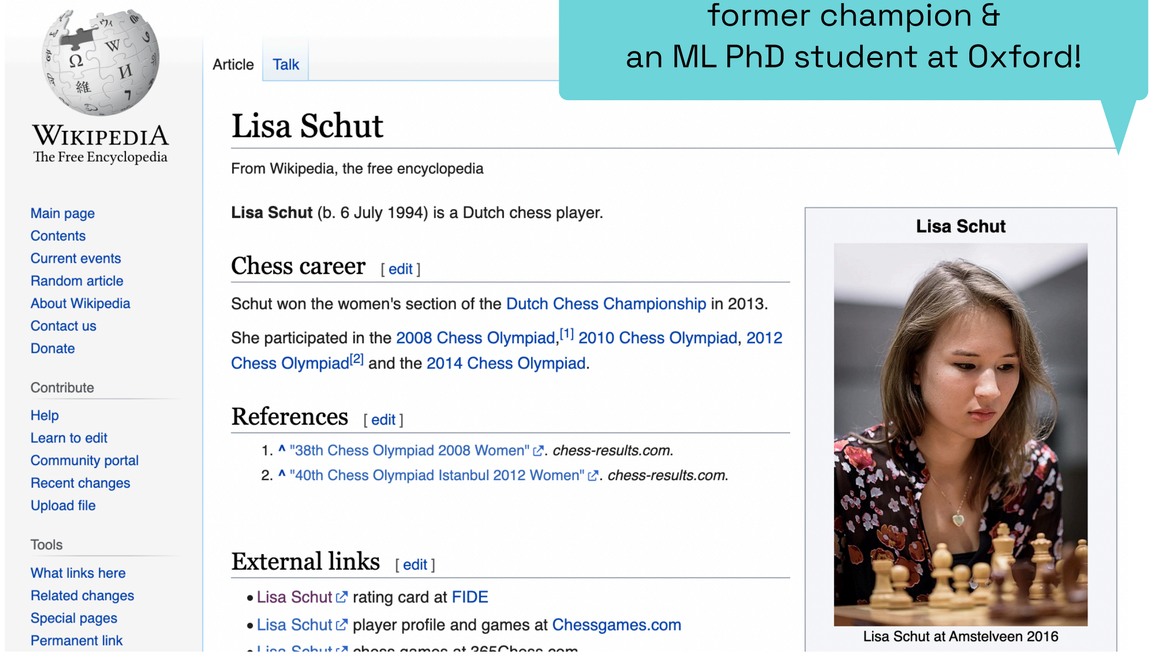


[ongoing] Teaching chess champions new super-human chess strategies

- Method: discover new chess strategies from AlphaZero by removing existing human concepts and leveraging new/existing concept's relationships.
- Evaluation: can the world chess champion (Magnus Carlsen) solve chess puzzles we generate to teach a new strategy?



This work is only possible because the first author of this work is a former champion & an ML PhD student at Oxford!



Small steps towards our hopes and dreams.

Improve tools we use to understand machines

1. Assumptions: built under wrong assumptions?
2. Expectations: not doing what we think they do?
3. Beyond us: humans can't understand them?

Observational study:
Given data, discover behavior

controlled study:
Intervene, observe, discover.