Topics:

- Bias/Fairness
- Wrap-up:
  - Open directions in Deep Learning
- Paper Discussion

CS 4644-DL / 7643-A
ZSOLT KIRA

- **Projects!**
  - Final project report due **July 26th**

- CIOS
  - Please make sure to fill out! Let us know about things you liked and didn’t like in comments so that we can keep or improve!
  - http://b.gatech.edu/cios

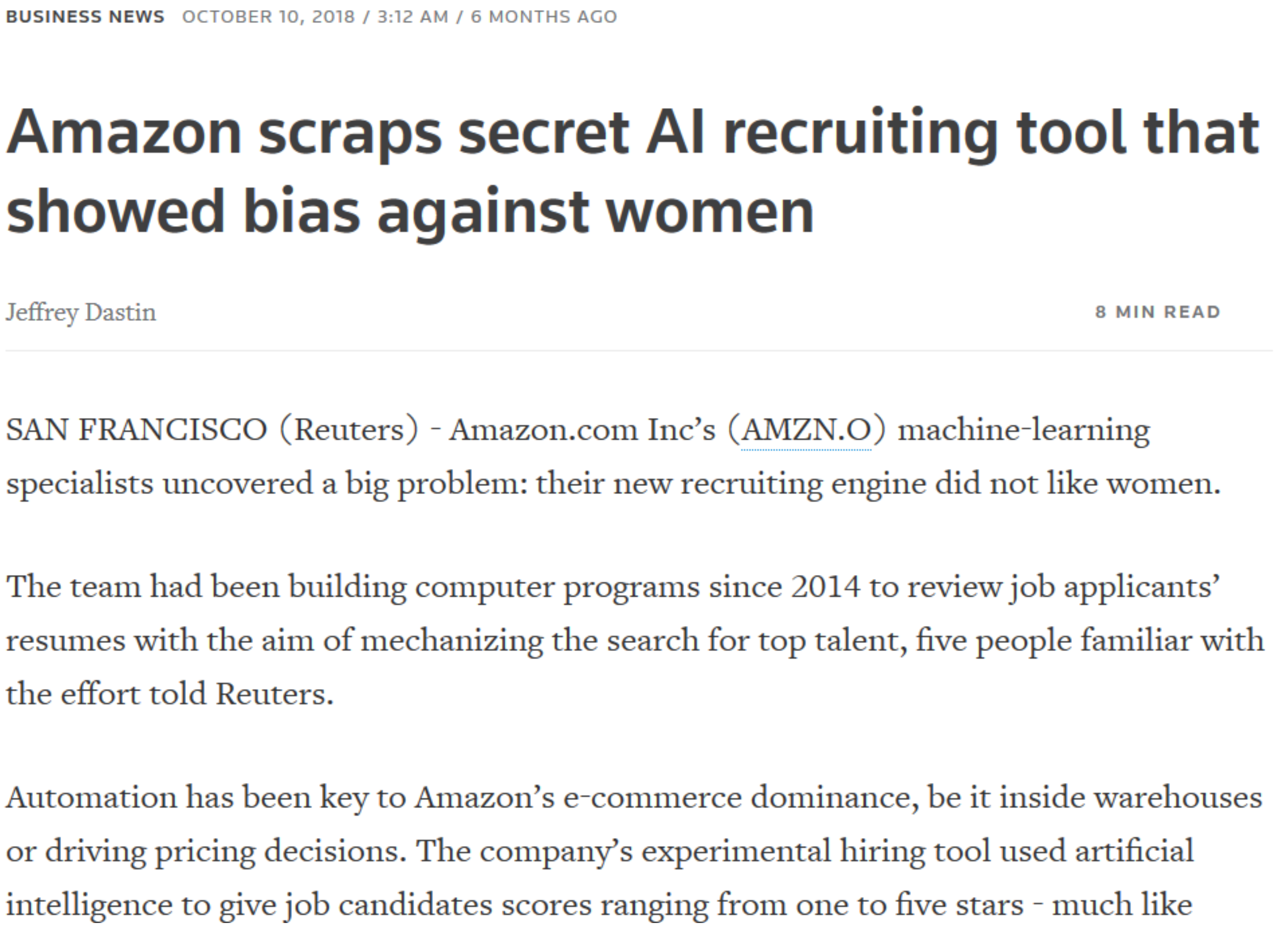
Bias & Fairness

ML and Fairness

- AI effects our lives in many ways
- Widespread algorithms with many small interactions
  - e.g. search, recommendations, social media
- Specialized algorithms with fewer but higher-stakes interactions
  - e.g. medicine, criminal justice, finance
- At this level of impact, algorithms can have unintended consequences
- Low classification error is not enough, need fairness

Slide By Aaron Roth

Slide Credit: David Madras



BUSINESS NEWS OCTOBER 10, 2018 / 3:12 AM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

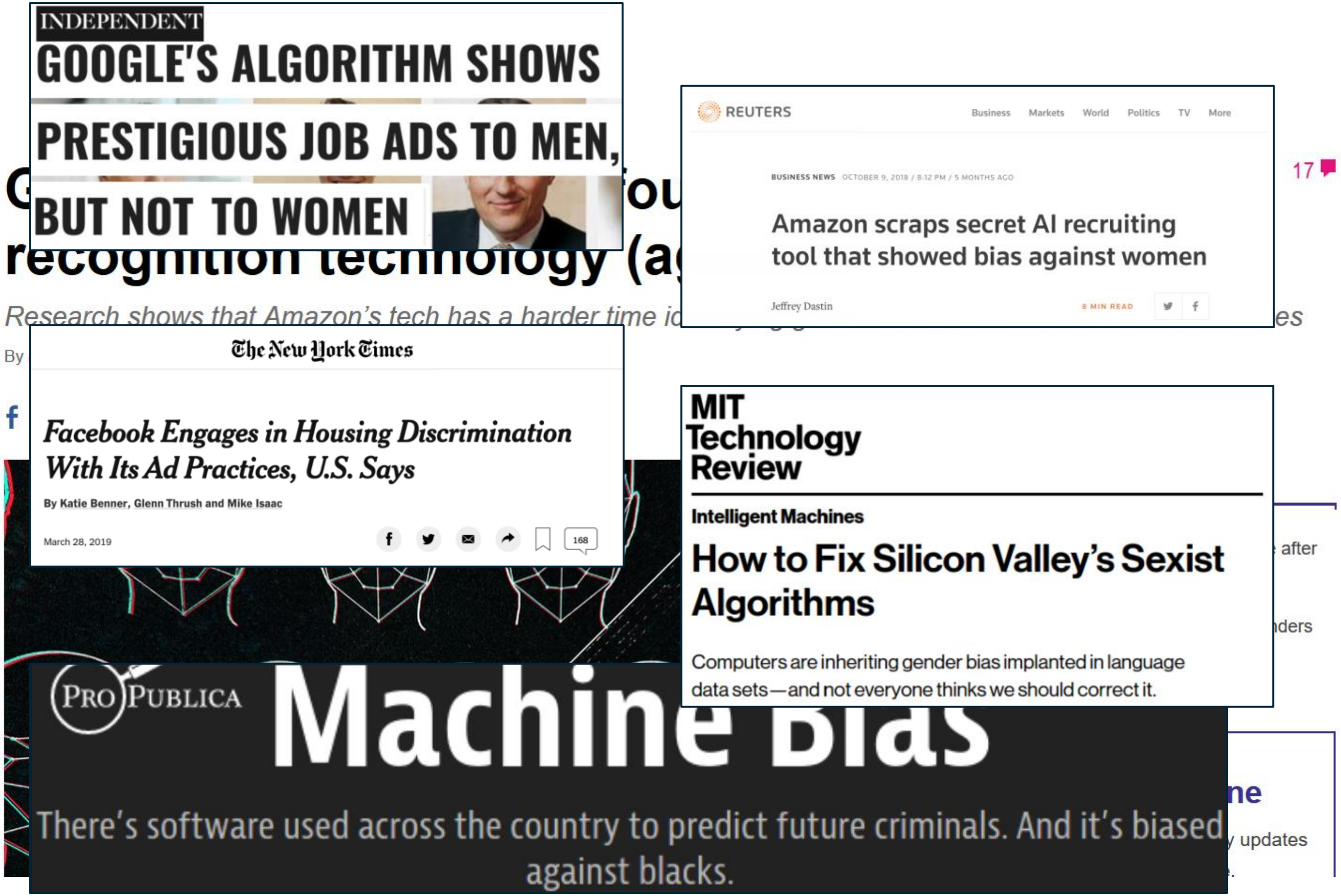
8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc’s (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants’ resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon’s e-commerce dominance, be it inside warehouses or driving pricing decisions. The company’s experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like

Slide By Aaron Roth



INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

G... fou... recognition technology (a...

Research shows that Amazon's tech has a harder time id... es

By

17

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

168

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

... after

...nders

...ne

...y updates

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Slide By Aaron Roth

Machine Learning and Social Norms

Fairness, Accountability, and Transparency in Machine Learning

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."

The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

- Sample norms: privacy, fairness, transparency, accountabılıty…
- Possible approaches
  - "traditional": legal, regulatory, watchdog
  - *Embed* social norms in data, algorithms, models
- Case study: privacy-preserving machine learning
  - "single", strong, definition (differential privacy)
  - almost every ML algorithm has a private version
- Fair machine learning
  - not so much…
  - impossibility results

Slide By Aaron Roth

(Un)Fairness Where?

- Data (input)
  - e.g. more arrests where there are more police
  - Label should be “committed a crime”, but is “convicted of a crime”
  - try to “correct” bias
- Models (output)
  - e.g. discriminatory treatment of subpopulations
  - build or “post-process” models with subpopulation guarantees
  - equality of false positive/negative rates; calibration
- Algorithms (process)
  - learning algorithm *generating* data through its decisions
  - e.g. don’t learn outcomes of denied mortgages
  - lack of clear train/test division
  - design (sequential) *algorithms* that are fair

Slide By Aaron Roth

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.

ImageNet (a very popular image dataset) with 1.2million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectely.

Data Bias

Slide By Hunter Schafer

Georgia Tech

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

- Financial responsibility → Credit Score
- Crime Rate → Arrest Rate
- Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren’t relevant to the task at hand), this can be another source of bias.

Measurement Bias

Slide By Hunter Schafer

Examples:

- If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.
  - Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.

- Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.
- The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

Measurement Bias

Slide By Hunter Schafer

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren’t allowed.
- Different definitions of fairness can be contradictory!

ML and Fairness

- Fairness is morally and legally motivated
- Takes many forms
- Criminal justice: recidivism algorithms (COMPAS)
  - Predicting if a defendant should receive bail
  - Unbalanced false positive rates: more likely to wrongly deny a black person bail

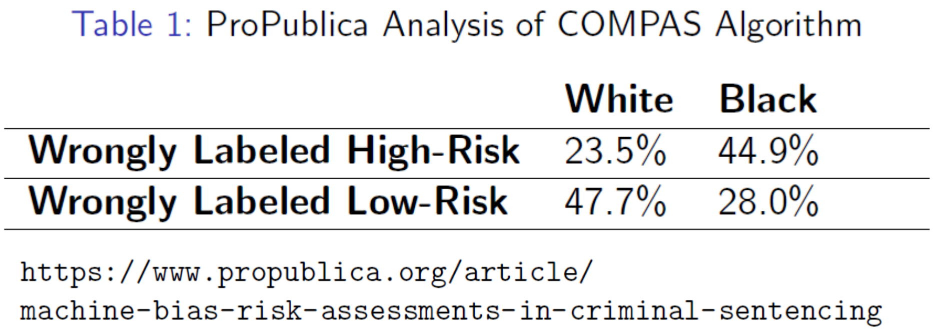


Table 1: ProPublica Analysis of COMPAS Algorithm

	White	Black
Wrongly Labeled High-Risk	23.5%	44.9%
Wrongly Labeled Low-Risk	47.7%	28.0%

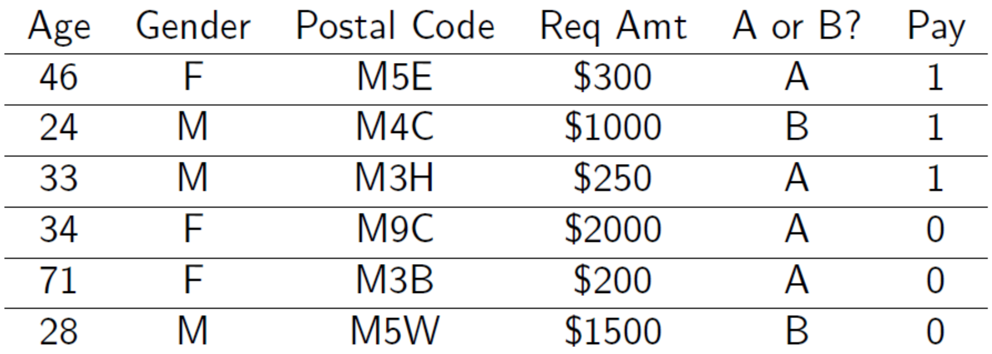
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Slide Credit: David Madras

Why Fairness is Hard

- Suppose we are a bank trying to fairly decide who should get a loan
  - i.e. Who is most likely to pay us back?
- Suppose we have two groups, A and B (the sensitive attribute)
  - This is where discrimination could occur
- The simplest approach is to remove the sensitive attribute from the data, so that our classier doesn't know the sensitive attribute. Often called **“Fairness through unawareness”**

Table 2: To Loan or Not to Loan?



Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	$300	A	1
24	M	M4C	$1000	B	1
33	M	M3H	$250	A	1
34	F	M9C	$2000	A	0
71	F	M3B	$200	A	0
28	M	M5W	$1500	B	0

Slide Credit: David Madras

Why Fairness is Hard

- However, if the sensitive attribute is correlated with the other attributes, this isn't good enough
- It is easy to predict race if you have lots of other information (e.g. home address, spending patterns)
- More advanced approaches are necessary

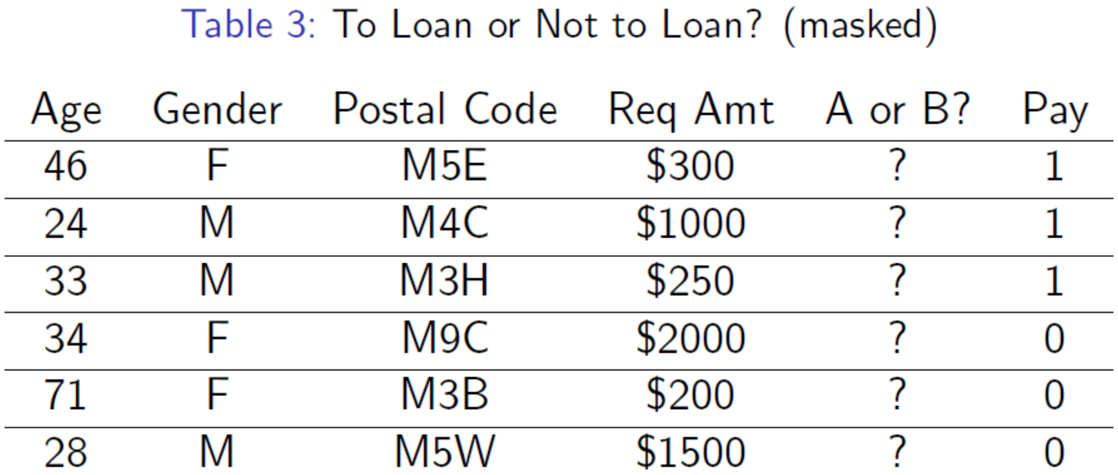


Table 3: To Loan or Not to Loan? (masked)

Age	Gender	Postal Code	Req Amt	A or B?	Pay
46	F	M5E	$300	?	1
24	M	M4C	$1000	?	1
33	M	M3H	$250	?	1
34	F	M9C	$2000	?	0
71	F	M3B	$200	?	0
28	M	M5W	$1500	?	0

Doesn’t work in practice. This does not prevent historical or measurement bias. Protected attributes can be unintentionally inferred from other, related attributes (e.g., in some cities, zip code can be deeply correlated with race).

Slide Credit: David Madras

Definitions of Fairness – Group Fairness

- So we've built our classier . . . how do we know if we're being fair?
- One metric is demographic parity | requiring that the same percentage of A and B receive loans
  - What if 80% of A is likely to repay, but only 60% of B is?
  - Then demographic parity is too strong
- Could require equal false positive/negative rates
  - When we make an error, the direction of that error is equally likely for both groups

$$P(loan|no\ repay, A) = P(loan|no\ repay, B)$$

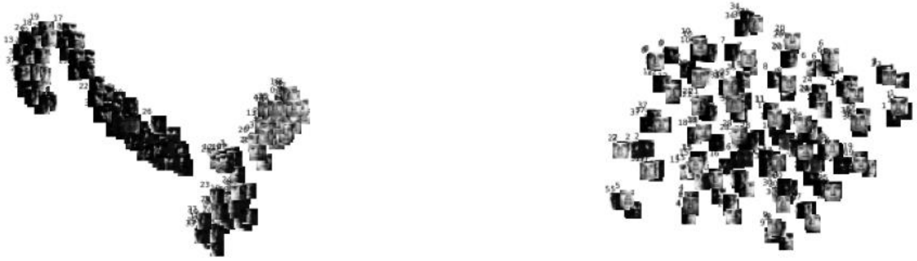
$$P(no\ loan|would\ repay, A) = P(no\ loan|would\ repay, B)$$

- These are definitions of group fairness
- Treat different groups equally"

Slide Credit: David Madras

Definitions of Fairness – Individual Fairness

- Also can talk about individual fairness | “Treat similar examples similarly"
- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaptation
  - Generative modelling/adversarial approaches



(a) Unfair representations

(b) Fair(er) representations

Figure 1: “The Variational Fair Autoencoder” (Louizos et al., 2016)

Slide Credit: David Madras

Conclusion

- This is an exciting field, quickly developing
- Central definitions still up in the air
- AI moves fast | lots of (currently unchecked) power
- Law/policy will one day catch up with technology
- Those who work with AI should be ready
  - **Think about implications of what you develop!**

Slide Credit: David Madras

Parting Thoughts

We Learned a Lot!

Deep Learning Fundamentals

Linear classification
Loss functions
Optimization
Optimizers
Backpropagation
Computation Graph
Multi-layer Perceptrons

Neural Network Components and Architectures

Hardware & software
Convolutions
Convolution Neural Networks
Pooling
Activation functions
Batch normalization
Transfer learning
Data augmentation
Architecture design
RNN/LSTMs
Attention & Transformers

Applications & Learning Algorithms

Semantic & instance Segmentation
Reinforcement Learning
Large-language Models
Variational Autoencoders
Diffusion Models
Generative Adversarial Nets
Self-supervised Learning
Vision-Language Models
VLM for Robotics

Some existing works not covered…

Current / Past

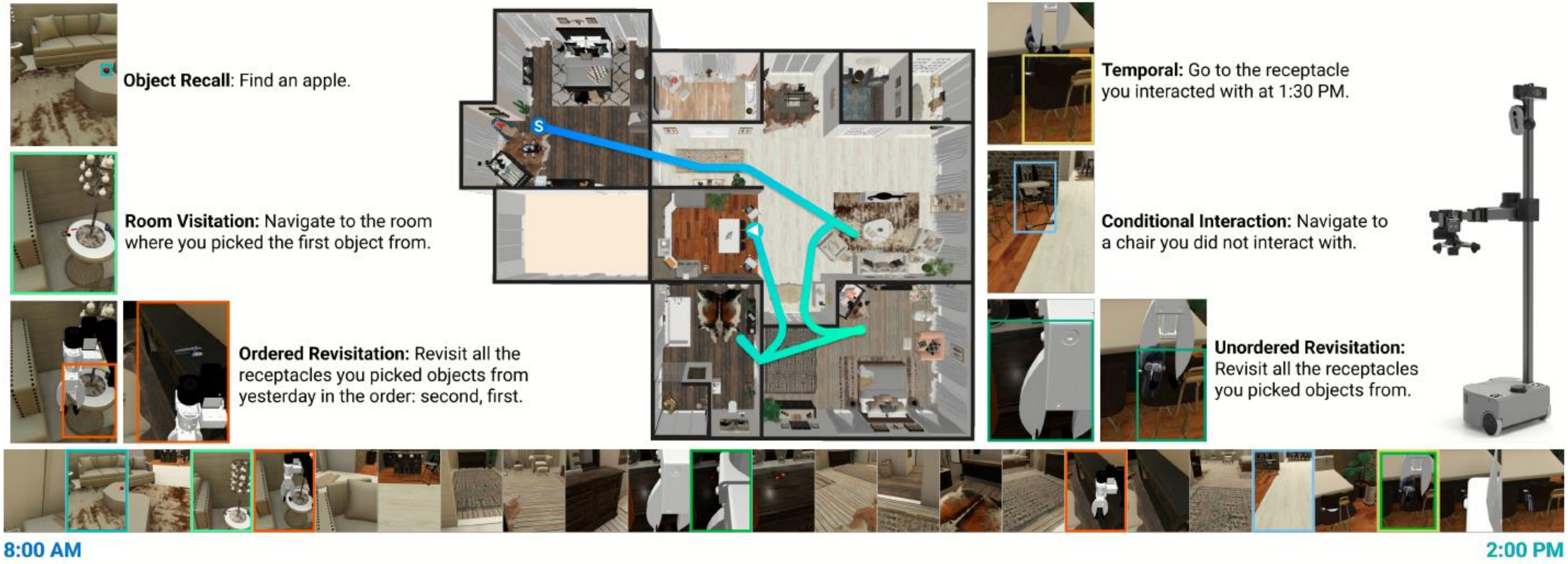
- Graph neural networks
- Meta-learning
- AutoML
- 3D perception & reconstruction / NeRFs
  - Neural Radiance Fields
- AI for Tabular data, time-series, etc.
- Beyond supervised learning: Semi-supervised, domain adaptation, zero/one/few-shot learning
- Embodied AI & Embodied question answering
- Adversarial Learning
- Continual/lifelong learning without forgetting
- World modeling, learning intuitive/physics models
- Reasoning, Planning, Search
- Neural Theorem Proving, induction & synthesis
- AI for science
- MLSys and MLOps
- Evaluation…
- Alignment
- Security

When Comparing to Humans, What’s Missing?

- Reasoning
  - ~~What does it mean for a neural network to “think” longer?~~
  - Chain-of-thought probably still off from how humans do it!
- Memory
- Planning, Search
- Deep integration of concepts and modalities
- Cognitive Architecture?

MA-LMM: Memory-Au
Model for Long-Ter

Spatially-Aware Transformer for Embodied Agents



Object Recall: Find an apple.

Room Visitation: Navigate to the room where you picked the first object from.

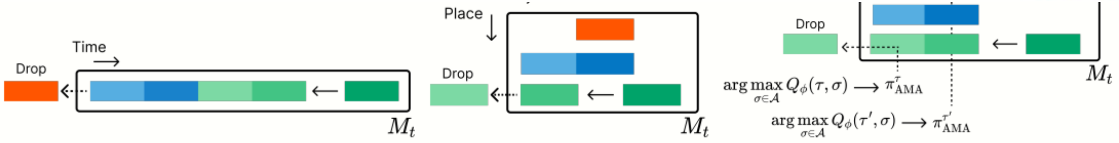
Ordered Revisitation: Revisit all the receptacles you picked objects from yesterday in the order: second, first.

Temporal: Go to the receptacle you interacted with at 1:30 PM.

Conditional Interaction: Navigate to a chair you did not interact with.

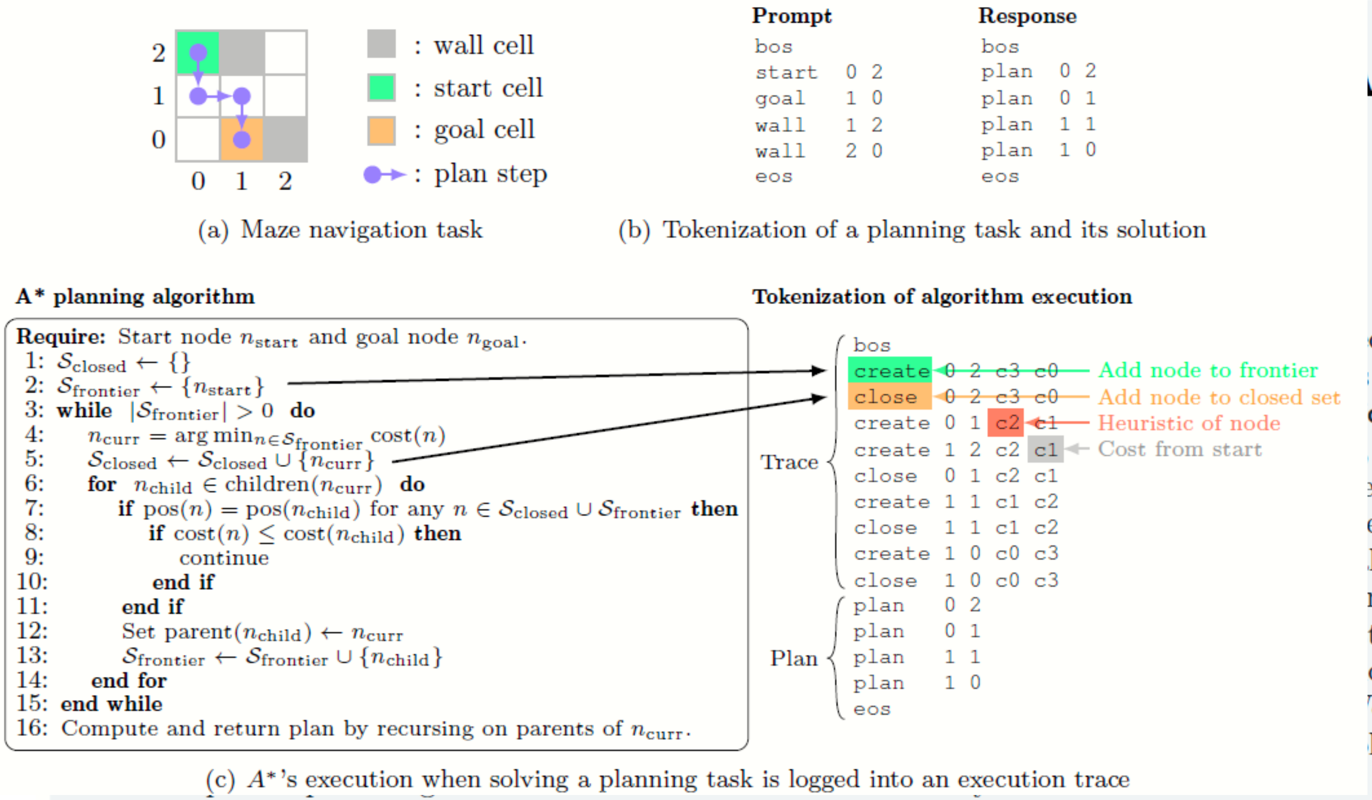
Unordered Revisitation: Revisit all the receptacles you picked objects from.

Graves et. al, Neural Turing Machines



Drop
Time →
M_t
Place
Drop
M_t
Drop
M_t
$\arg\max_{\sigma \in \mathcal{A}} Q_\phi(\tau, \sigma) \longrightarrow \pi^{\tau}_{\text{AMA}}$
$\arg\max_{\sigma \in \mathcal{A}} Q_\phi(\tau', \sigma) \longrightarrow \pi^{\tau'}_{\text{AMA}}$

(a) FIFO Memory **(b)** Place Memory **(c)** Adaptive Memory Allocator (AMA)



(a) Maze navigation task

Prompt

```
bos
start   0 2
goal    1 0
wall    1 2
wall    2 0
eos
```

Response

```
bos
plan   0 2
plan   0 1
plan   1 1
plan   1 0
eos
```

(b) Tokenization of a planning task and its solution

A* planning algorithm

```
Require: Start node $n_{\text{start}}$ and goal node $n_{\text{goal}}$.
$\mathcal{S}_{\text{closed}} \leftarrow \{\}$
$\mathcal{S}_{\text{frontier}} \leftarrow \{n_{\text{start}}\}$
while $|\mathcal{S}_{\text{frontier}}| > 0$ do
    $n_{\text{curr}} = \arg\min_{n \in \mathcal{S}_{\text{frontier}}} \text{cost}(n)$
    $\mathcal{S}_{\text{closed}} \leftarrow \mathcal{S}_{\text{closed}} \cup \{n_{\text{curr}}\}$
    for $n_{\text{child}} \in \text{children}(n_{\text{curr}})$ do
        if $\text{pos}(n) = \text{pos}(n_{\text{child}})$ for any $n \in \mathcal{S}_{\text{closed}} \cup \mathcal{S}_{\text{frontier}}$ then
            if $\text{cost}(n) \leq \text{cost}(n_{\text{child}})$ then
                continue
            end if
        end if
        Set $\text{parent}(n_{\text{child}}) \leftarrow n_{\text{curr}}$
        $\mathcal{S}_{\text{frontier}} \leftarrow \mathcal{S}_{\text{frontier}} \cup \{n_{\text{child}}\}$
    end for
end while
Compute and return plan by recursing on parents of $n_{\text{curr}}$.
```

Tokenization of algorithm execution

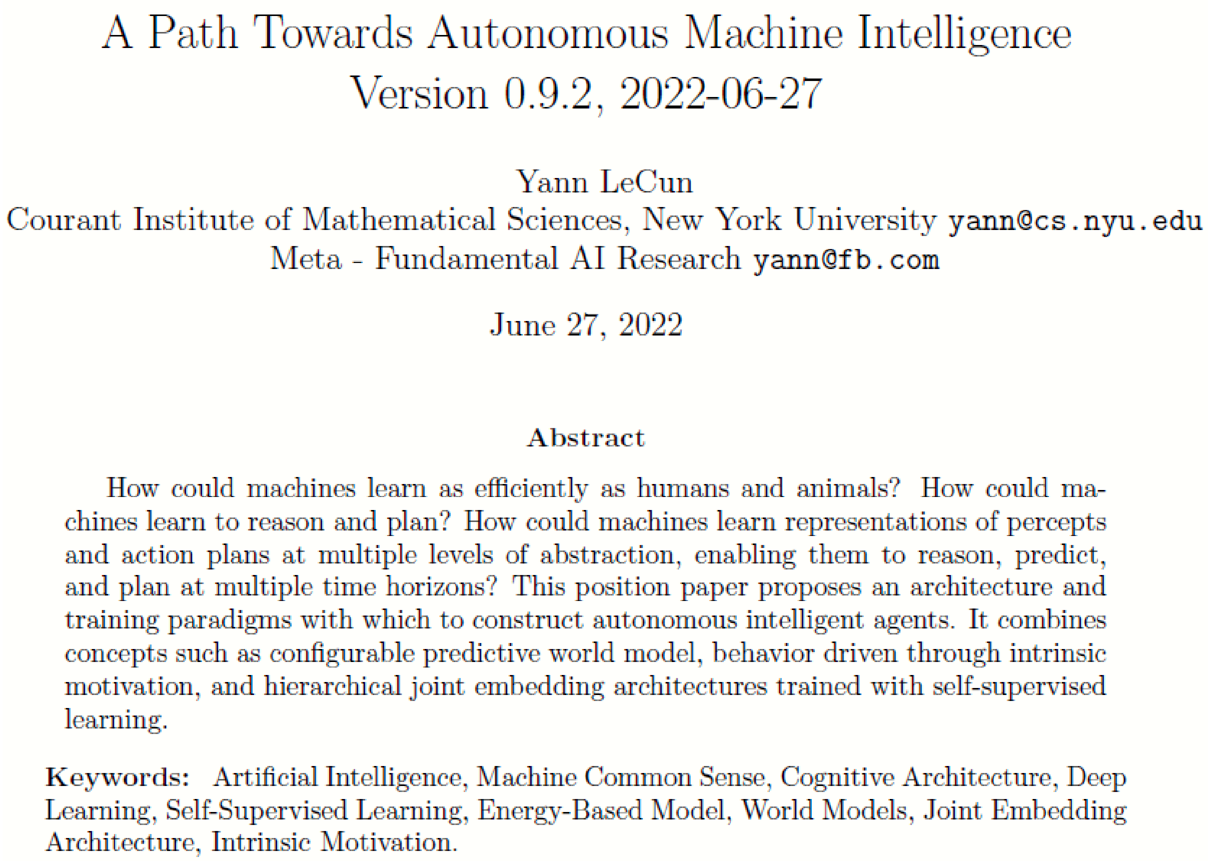
```
        bos
        create  0 2 c3 c0   Add node to frontier
        close   0 2 c3 c0   Add node to closed set
        create  0 1 c2 c1   Heuristic of node
        create  1 2 c2 c1   Cost from start
Trace   close   0 1 c2 c1
        create  1 1 c1 c2
        close   1 1 c1 c2
        create  1 0 c0 c3
        close   1 0 c0 c3
        plan    0 2
        plan    0 1
Plan    plan    1 1
        plan    1 0
        eos
```

(c) *A**'s execution when solving a planning task is logged into an execution trace

via

ctures
work,
ormer,
while
ecoder
ed via
l plan.
n task
tudies
ct the
'e also
koban

Correspondence: {lucaslehnert, yuandong}@meta.com



A Path Towards Autonomous Machine Intelligence
Version 0.9.2, 2022-06-27

Yann LeCun
Courant Institute of Mathematical Sciences, New York University yann@cs.nyu.edu
Meta - Fundamental AI Research yann@fb.com

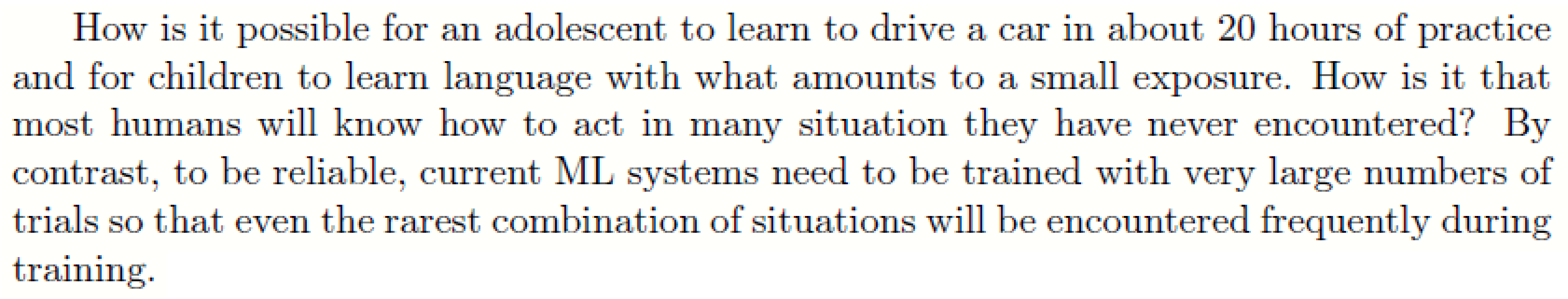
June 27, 2022

Abstract

How could machines learn as efficiently as humans and animals? How could machines learn to reason and plan? How could machines learn representations of percepts and action plans at multiple levels of abstraction, enabling them to reason, predict, and plan at multiple time horizons? This position paper proposes an architecture and training paradigms with which to construct autonomous intelligent agents. It combines concepts such as configurable predictive world model, behavior driven through intrinsic motivation, and hierarchical joint embedding architectures trained with self-supervised learning.

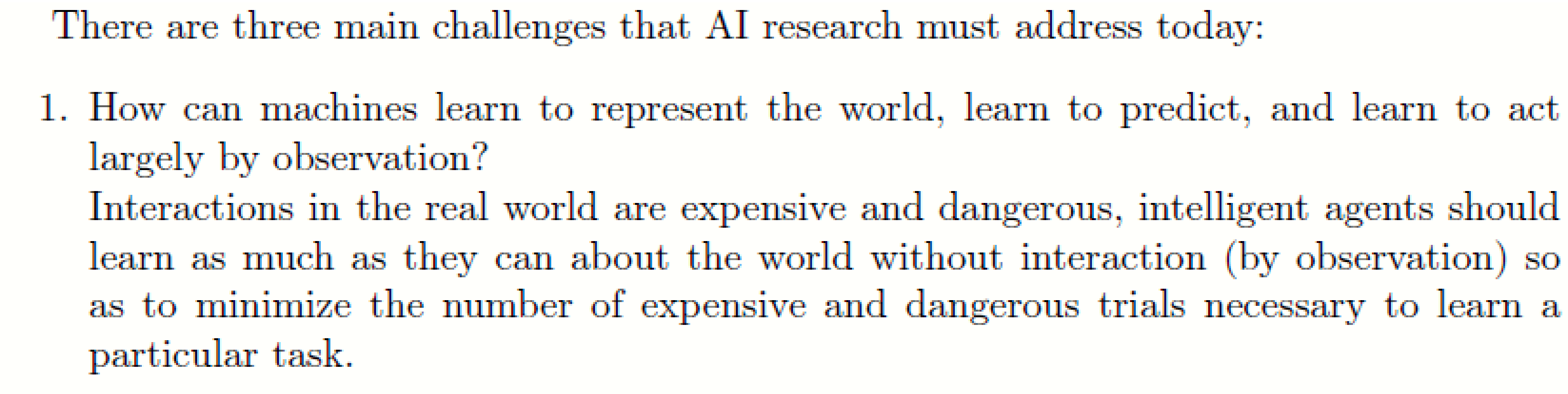
Keywords: Artificial Intelligence, Machine Common Sense, Cognitive Architecture, Deep Learning, Self-Supervised Learning, Energy-Based Model, World Models, Joint Embedding Architecture, Intrinsic Motivation.

Motivation



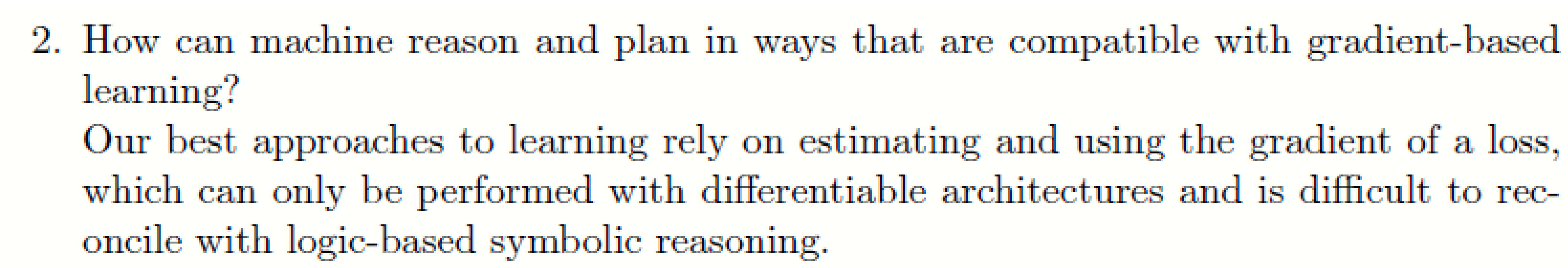
How is it possible for an adolescent to learn to drive a car in about 20 hours of practice and for children to learn language with what amounts to a small exposure. How is it that most humans will know how to act in many situation they have never encountered? By contrast, to be reliable, current ML systems need to be trained with very large numbers of trials so that even the rarest combination of situations will be encountered frequently during training.

Challenges

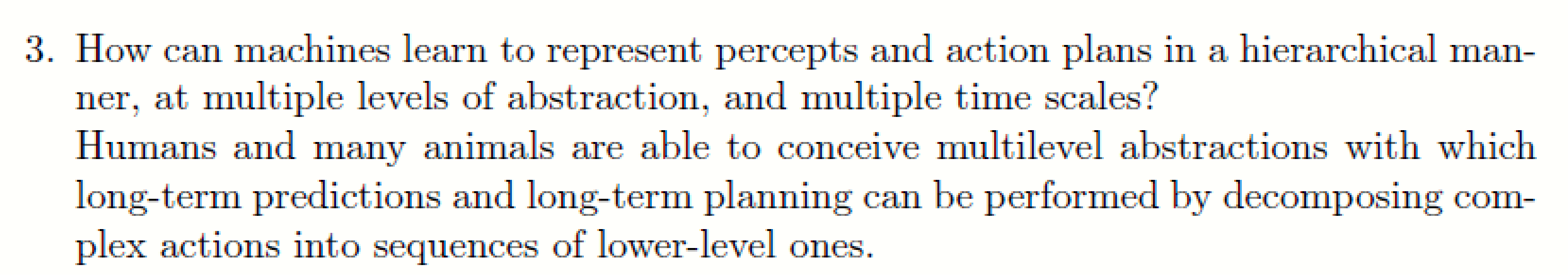


There are three main challenges that AI research must address today:

1. How can machines learn to represent the world, learn to predict, and learn to act largely by observation?
   Interactions in the real world are expensive and dangerous, intelligent agents should learn as much as they can about the world without interaction (by observation) so as to minimize the number of expensive and dangerous trials necessary to learn a particular task.



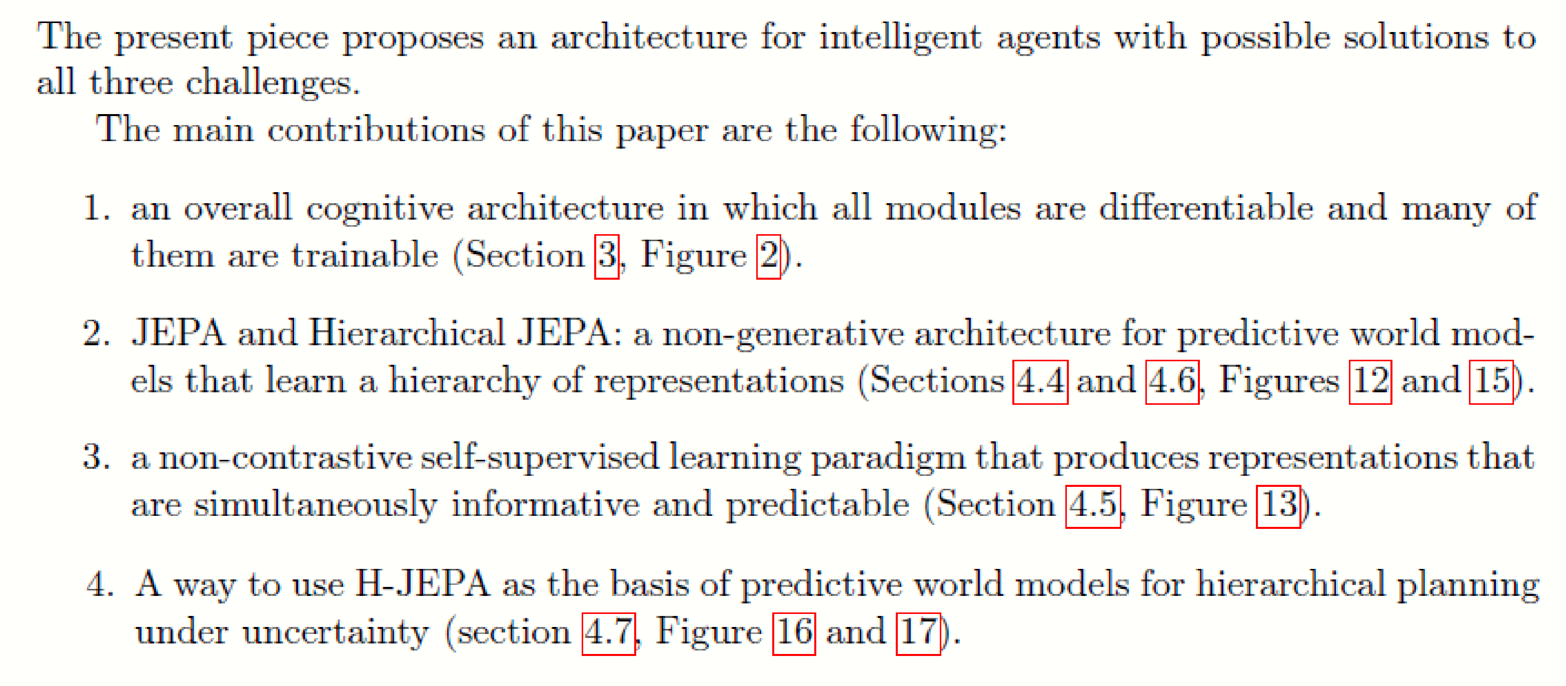
2. How can machine reason and plan in ways that are compatible with gradient-based learning?
   Our best approaches to learning rely on estimating and using the gradient of a loss, which can only be performed with differentiable architectures and is difficult to reconcile with logic-based symbolic reasoning.



3. How can machines learn to represent percepts and action plans in a hierarchical manner, at multiple levels of abstraction, and multiple time scales?
   Humans and many animals are able to conceive multilevel abstractions with which long-term predictions and long-term planning can be performed by decomposing complex actions into sequences of lower-level ones.

Are these the right challenges? Anything else?

Contributions



The present piece proposes an architecture for intelligent agents with possible solutions to all three challenges.

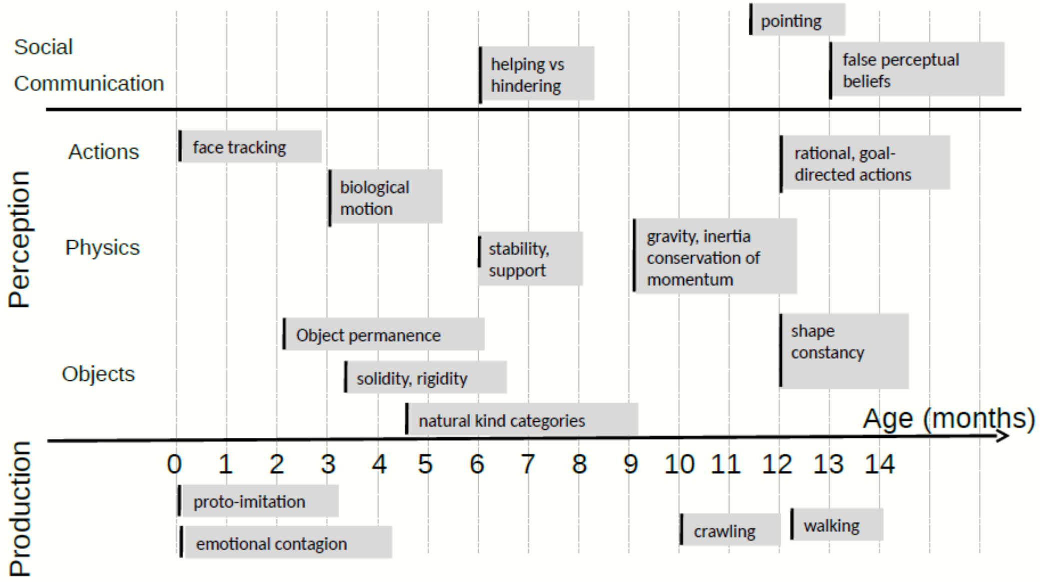
The main contributions of this paper are the following:

1. an overall cognitive architecture in which all modules are differentiable and many of them are trainable (Section 3, Figure 2).
2. JEPA and Hierarchical JEPA: a non-generative architecture for predictive world models that learn a hierarchy of representations (Sections 4.4 and 4.6, Figures 12 and 15).
3. a non-contrastive self-supervised learning paradigm that produces representations that are simultaneously informative and predictable (Section 4.5, Figure 13).
4. A way to use H-JEPA as the basis of predictive world models for hierarchical planning under uncertainty (section 4.7, Figure 16 and 17).

World Models

- Common sense can be seen as a collection of models of the world that can tell an agent what is likely, what is plausible, and what is impossible. Using such world models, animals can learn new skills with very few trials. They can predict the consequences of their actions, they can reason, plan, explore, and imagine new solutions to problems. Importantly, they can also avoid making dangerous mistakes when facing an unknown situation.

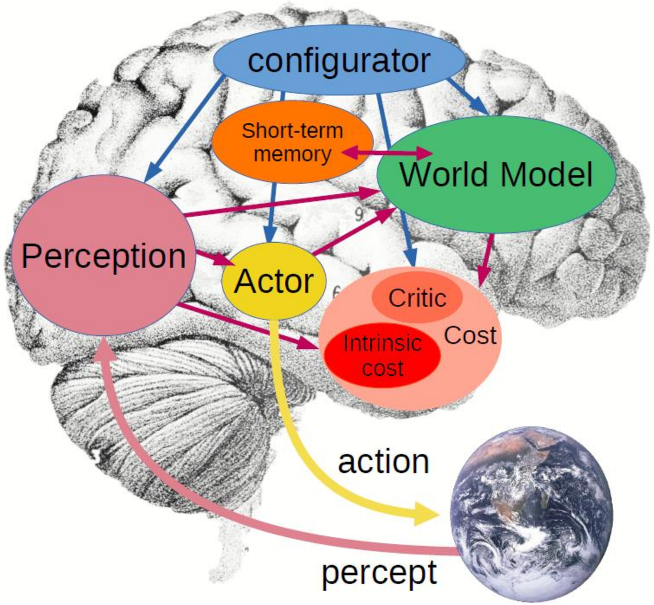
Humans & World Models



But can a human or animal brain contain all the world models that are necessary for survival? One hypothesis in this paper is that animals and humans have only one world model engine somewhere in their prefrontal cortex. That world model engine is dynamically configurable for the task at hand.

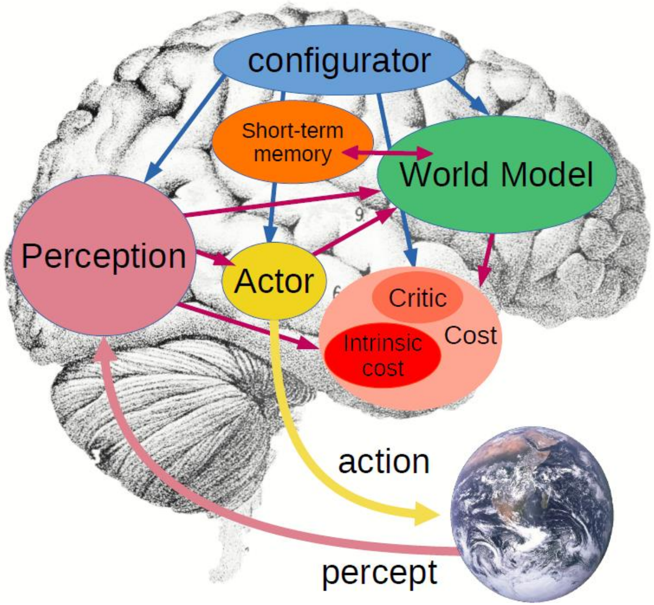
Architecture?

- The **configurator** module takes input from all other modules and configures them for the task at hand by modulating their parameters and their attention circuits.
- The **perception module** receives signals from sensors and estimates the current state of the world.
- The **world model module** constitutes the most complex piece of the architecture. Its role is twofold: (1) estimate missing information about the state of the world not provided by perception, (2) predict plausible future states of the world.



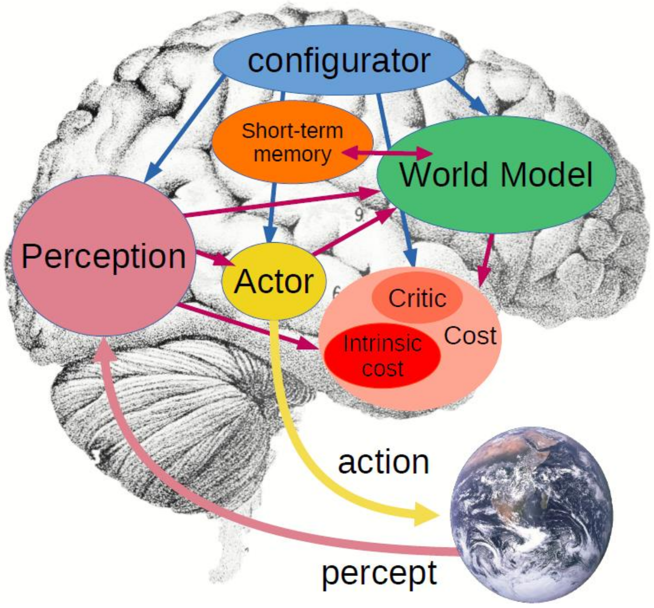
Architecture?

- The **cost module** measures the level of “discomfort" of the agent, in the form of a scalar quantity called the energy. The energy is the sum of two energy terms computed by two sub-modules: the Intrinsic Cost module and the Trainable Critic module. The overall objective of the agent is to take actions so as to remain in states that minimize the average energy.



Architecture?

- The **short-term memory module** stores relevant information about the past, current, and future states of the world, as well as the corresponding value of the intrinsic cost.
- The **actor module** computes proposals for sequences of actions and outputs actions to the eectors.



Mode 1

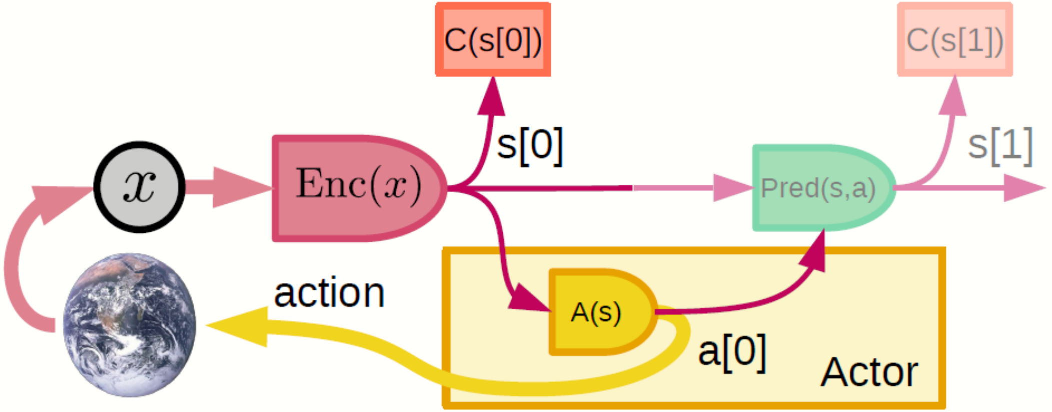
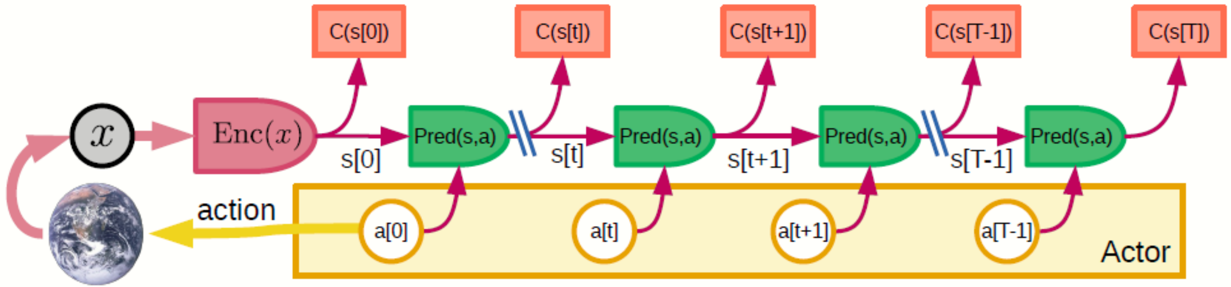


Figure 3: *Mode-1 perception-action episode. The perception module estimates the state of the world* $s[0] = \text{Enc}(x)$*. The actor directly computes an action, or a short sequence of actions, through a policy module* $a[0] = A(s[0])$*.*

This reactive process does not make use of the world model nor of the cost. The cost module computes the energy of the initial state $f[0] = \text{C}(s[0])$ *and stores the pairs* $(s[0], f[0])$ *in the short-term memory. Optionally, it may also predict the next state using the world model* $s[1] = \text{Pred}(s[0], a[0])$*, and the associated energy* $f[0] = \text{C}(s[0])$ *so that the world model can be adjusted once the next observation resulting from the action taken becomes available.*

Mode 2



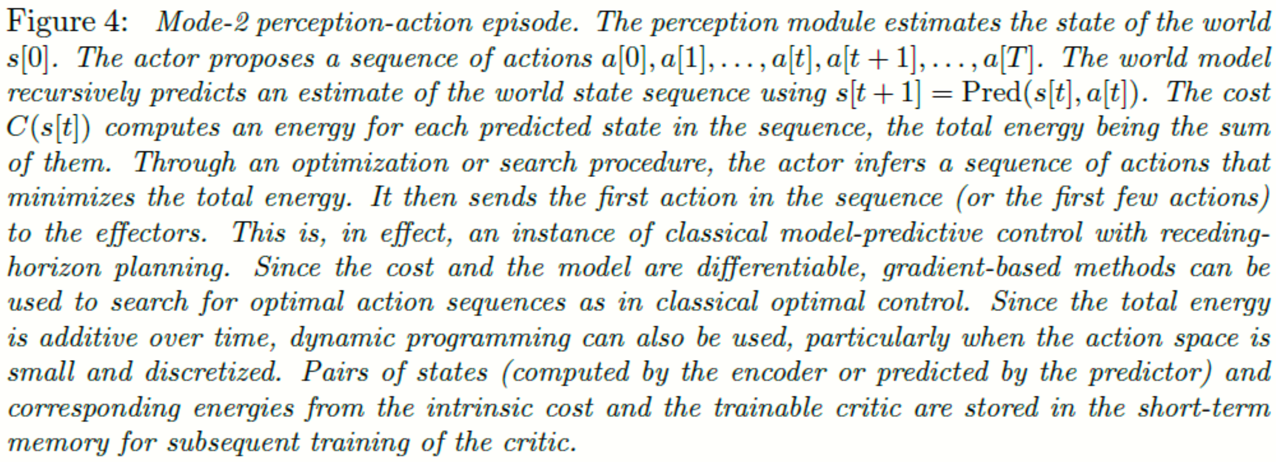


Figure 4: *Mode-2 perception-action episode. The perception module estimates the state of the world* $s[0]$*. The actor proposes a sequence of actions* $a[0], a[1], \ldots, a[t], a[t+1], \ldots, a[T]$*. The world model recursively predicts an estimate of the world state sequence using* $s[t+1] = \text{Pred}(s[t], a[t])$*. The cost* $C(s[t])$ *computes an energy for each predicted state in the sequence, the total energy being the sum of them. Through an optimization or search procedure, the actor infers a sequence of actions that minimizes the total energy. It then sends the first action in the sequence (or the first few actions) to the effectors. This is, in effect, an instance of classical model-predictive control with receding-horizon planning. Since the cost and the model are differentiable, gradient-based methods can be used to search for optimal action sequences as in classical optimal control. Since the total energy is additive over time, dynamic programming can also be used, particularly when the action space is small and discretized. Pairs of states (computed by the encoder or predicted by the predictor) and corresponding energies from the intrinsic cost and the trainable critic are stored in the short-term memory for subsequent training of the critic.*

Energy-Based Models?

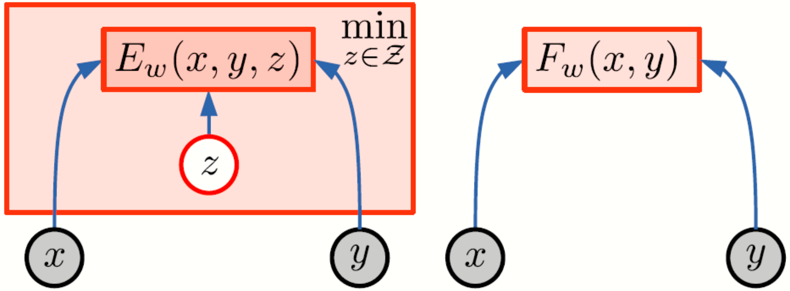
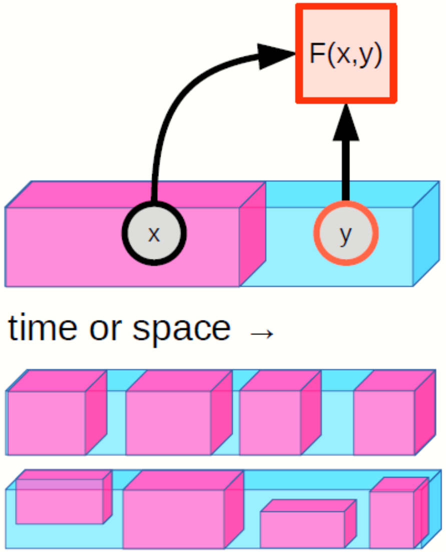
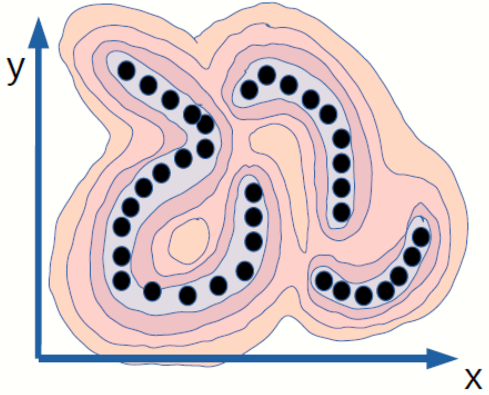


Figure 9: *Latent-Variable Energy-Based Model (LVEBM).* *To evaluate the degree of compatibility between x and y, an EBM may need the help of a latent variable z. The latent variable can be seen as parameterizing the set of possible relationships between an x and a set of compatible y. Latent variables represent information about y that cannot be extracted from x. For example, if x is a view of an object, and y another view of the same object, z may parameterize the camera displacement between the two views. Inference consists in finding the latent that minimizes the energy $\check{z} = \text{argmin}_{z \in \mathcal{Z}} E_w(x,y,z)$. The resulting energy $F_w(x,y) = E_w(x,y,\check{z})$ only depends on x and y. In the dual view example, inference finds the camera motion that best explains how x could be transformed into y.*



Energy Function



JEPA?

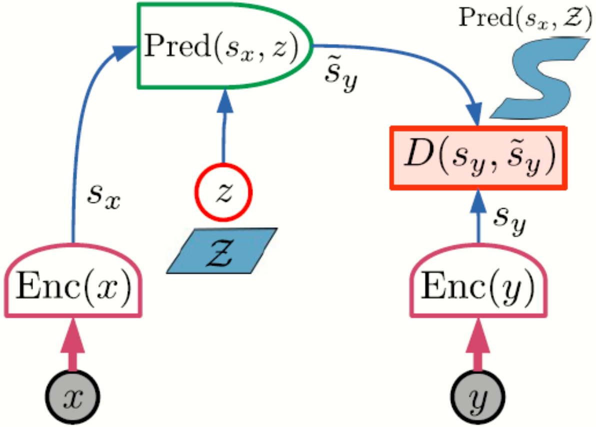


Figure 12: *The Joint-Embedding Predictive Architecture (JEPA) consists of two encoding branches. The first branch computes s_x, a representation of x and the second branch s_y a representation of y. The encoders do not need to be identical. A predictor module predicts s_y from s_x with the possible help of a latent variable z. The energy is the prediction error. Simple variations of the JEPA may use no predictor, forcing the two representations to be equal, or may use a fixed predictor with no latent, or may use simple latents such as discrete variables.*

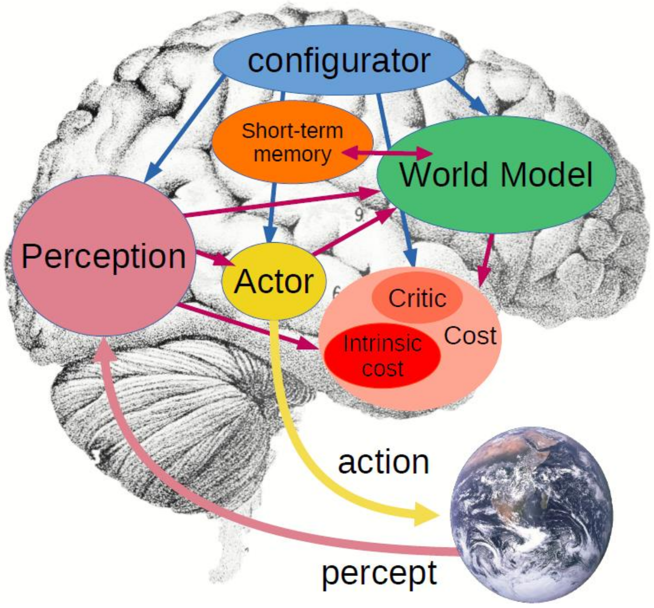
The main advantage of JEPA is that it performs predictions in representation space, eschewing the need to predict every detail of y, and enabling the elimination of irrelevant details by the encoders.

More precisely, the main advantage of this architecture for representing multi-modal dependencies is twofold: (1) the encoder function $s_y = \text{Enc}(y)$ may possess invariance properties that will make it produce the same s_y for a set of different y. This makes the energy constant over this set and allows the model to capture complex multi-modal dependencies; (2) The latent variable z, when varied over a set $\mathcal{Z}$, can produce a set of plausible predictions $\text{Pred}(s_x, \mathcal{Z}) = \{\tilde{s}_y = \text{Pred}(s_x, z)\ \forall z \in \mathcal{Z}\}$

If x is a video clip of a car approaching a fork in the road, s_x and s_y may represent the position, orientation, velocity and other characteristics of the car before and after the fork, respectively, ignoring irrelevant details such as the trees bordering the road or the texture of the sidewalk. z may represent whether the car takes the left branch or the right branch of the road.

Some Concepts

- World models: Predict missing data/future
- Configurability
- Intrinsic reward
- Prediction of reward (critic)
- Memory (short and long-term
- System 1 vs. System 2
- Embodiment



Discussions

- Do we need world models? Should they be explicit or implicit?
  - "Arguably, designing architectures and training paradigms for the world model constitute the main obstacles towards real progress in AI over the next decades."
- Do world models cover everything?
  - Math/reasoning, science/innovation, etc.
- Is reasoning now solved with inference-time methods?
- Do we need explicit symbols / symbolic reasoning?
- Does this get you robustness?
- Do we need embodiment to "solve" intelligence?

Open Discussion

Things to Watch out For

- Research is cyclical
  - SVMs, boosting, probabilistic graphical models & Bayes Nets, Structural Learning, Sparse Coding, Deep Learning
  - Deep learning is unique in its depth and breadth, but...
  - Deep learning may be improved, reinvented, combined, overtaken

- Learn fundamentals for techniques across the field:
  - Know the span of ML techniques and choose the ones that fit your problem!
  - **Be responsible** in 1) how you use it, 2) promises you make and how you convey it

- Try to understand landscape of the field
  - Look out for what is coming up next, not where we are

- Have fun!

Thank you!