ILIAD Intensive

Computational Mechanics

1. Overview & Scope

Xavier Poncini (Simplex)

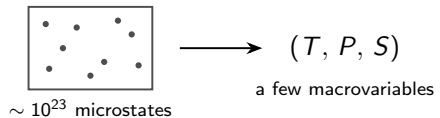
Autumn 2026

Outline

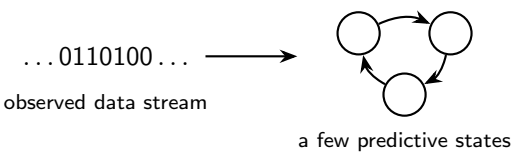
1. What is Computational Mechanics?
2. Computational Mechanics & AI Safety
3. Overview: where is the program at?
4. Scope: what will we tackle today?

What is Computational Mechanics?

Statistical Mechanics makes predictions about the behaviour of systems by considering their **statistical** properties:



Computational Mechanics makes predictions about the behaviour of systems by considering their **computational** properties:



Central questions

Some questions that are central to comp-mech:

- How much of the past does a system store?
- How does the system store this information?
- How does this information affect system behaviour?

We will consider each of these questions for transformers today!

Optimal prediction

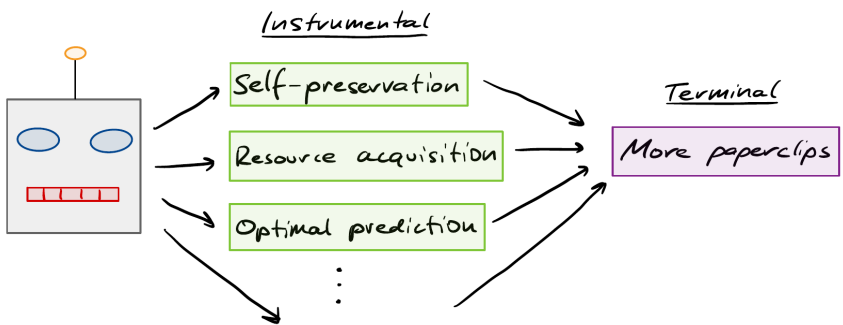
Not all systems are interesting! Comp-mech places an emphasis on those performing **optimal prediction**. In this setting the central question becomes:

> What are convergent structures relevant for **optimal prediction**?

Insight on this question greatly constrains our search space when considering **optimal predictors**.

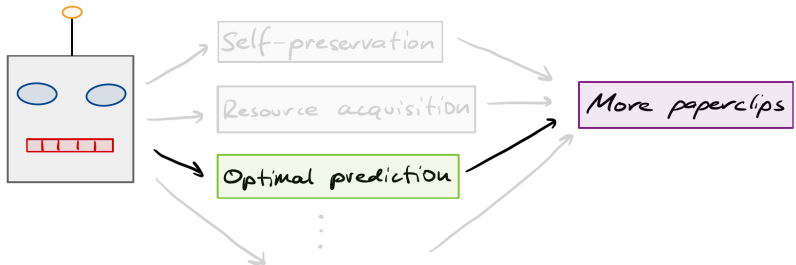
Instrumental and terminal goals

We have strong reason to believe that AI systems will pursue a range of **instrumental goals** in service of their **terminal goal**.



Optimal prediction as an intervention target

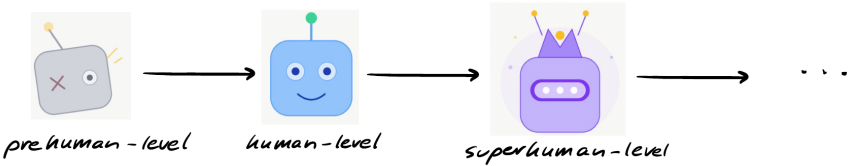
Comp-mech uses an **instrumental goal** as an intervention target point.



By considering internal representations consistent with optimal prediction the search space is reduced.

Capability scale and optimal prediction

Targeting an instrumental goal is **scale-free**: it will be relevant to current and future models.



Which contrasts **scale-sensitive** approaches e.g. AI control that are only applicable up to a fixed capability scale.

Moreover, as capabilities increase models will better approximate **optimal predictors**.

The research program

Comp-mech applied to AI safety is a young research program (∼2.5 years). There is much to do!

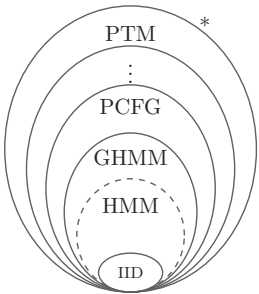
Progress can be broken down into two orthogonal directions:

- For what classes of data generating **processes** do we understand optimal prediction?
- For what scale of **neural networks** have we identified the structures of optimal prediction?

Processes and generation resources

Fix a (computable) stochastic process and consider:

What resources are required to generate this process? This naturally leads to a hierarchy:

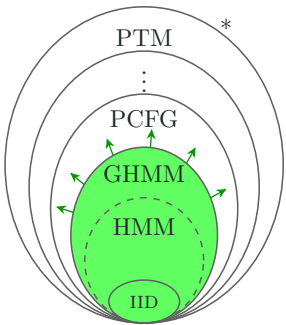


∗ not exhaustive

PTM	Probabilistic Turing Machine
PCFG	Probabilistic Context-Free Grammar
GHMM	Generalised Hidden Markov Model
HMM	Hidden Markov Model
IID	Independent & Identically Distributed

Processes and optimal prediction

For what classes of data generating **processes** do we understand optimal prediction?



∗ not exhaustive

PTM	Probabilistic Turing Machine
PCFG	Probabilistic Context-Free Grammar
GHMM	Generalised Hidden Markov Model
HMM	Hidden Markov Model
IID	Independent & Identically Distributed

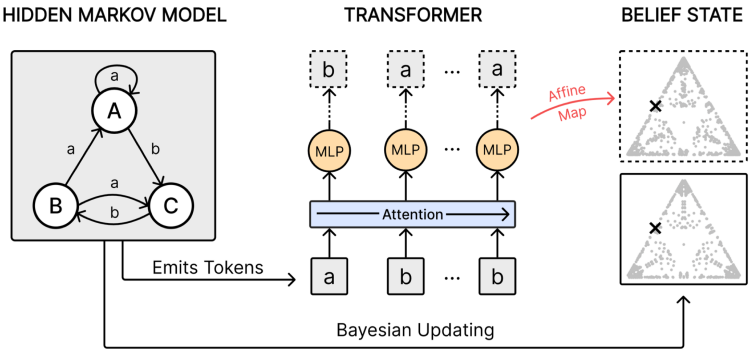
Neural networks: the scale studied so far

For what scale of **neural networks** have we identified the structures of optimal prediction?

Architecture	Parameters (approximately)
Transformer	10^5
LSTM	10^5
RNN	10^4
GRU	10^5

Very toy compared to current frontier models ($\sim 10^{12}$)!

Example: belief state geometry

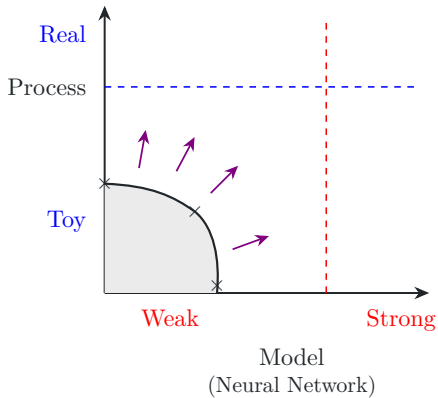


After next-token training on HMM sequences, an affine readout approximately recovers the geometry of beliefs over hidden states from residual activations.

Shai et al., *Transformers represent belief state geometry in their residual stream.* NeurIPS 2024. arXiv:2405.15943.

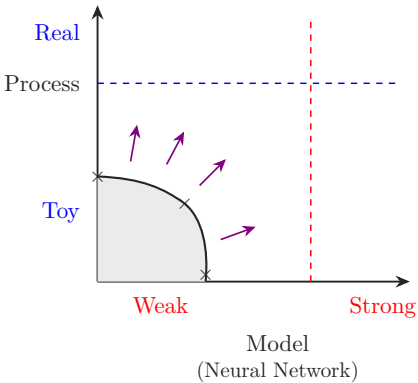
The research frontier

We are working to push the Pareto frontier:



The research frontier

We are working to push the Pareto frontier:

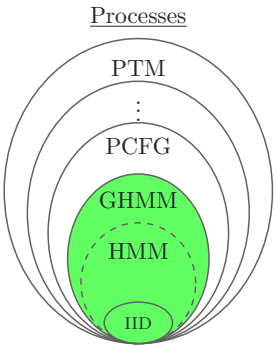


Wild Speculation

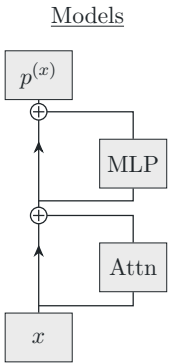
As the research program is scale-free one view is that we should develop the foundations such that future automated AI Safety researchers can pursue these ideas independently.

Scope: what will we tackle today?

We will take a slightly narrower view and consider:



GHMMs



Transformers

Plan

1	Predicting the future	Lecture/Exercises
2	Representing the past	Lecture/Exercises
3	Belief geometry in transformers	Reading/Discussion

Questions?