Computational Mechanics:

1. Overview & Scope

ICIAD INTENSIVE – SPRING 2026

Outline:

1) What is Computational Mechanics?

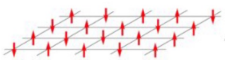
2) Computational Mechanics & AI Safety

3) Overview – where is the program at?

4) Scope – what will we tackle today?

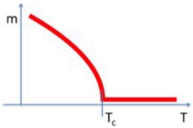
What is Computational Mechanics?

* Statistical mechanics makes predictions about the behaviour of systems by considering their statistical properties. E.g.



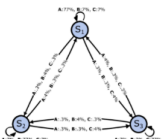
predicts

ferromagnetism



Figures from Yuvan et al.

* Computational mechanics (comp-mech) makes predictions about the behaviour of systems by considering their computational properties. E.g.



predicts

structure in activations

Figures from Shai et al.

What is Computational Mechanics?

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!

What is Computational Mechanics?

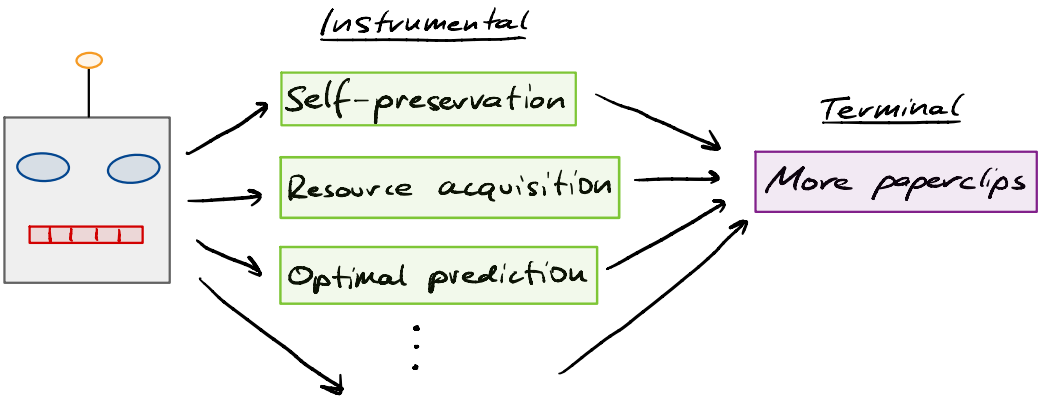
Not all systems are interesting! Comp-mech places an emphasis on those performing optimal prediction. In this setting the central question becomes:

> What are convergent structures relevant for optimal prediction?

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

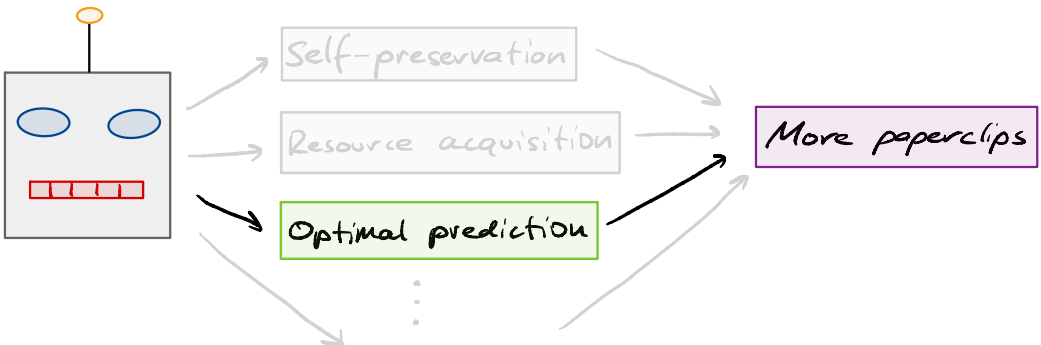
Computational Mechanics & AI Safety

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



Computational Mechanics & AI Safety

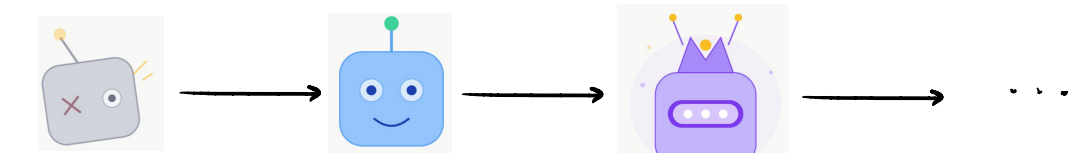
Comp-mech uses an instrumental goal as an intervention target point



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

Computational Mechanics & AI Safety

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.

Overview — where is the program at?

Comp-mech applied to AI safety is a young research program (~ 2 years). There is much to do!

Progress can be broken down into two orthogonal directions:

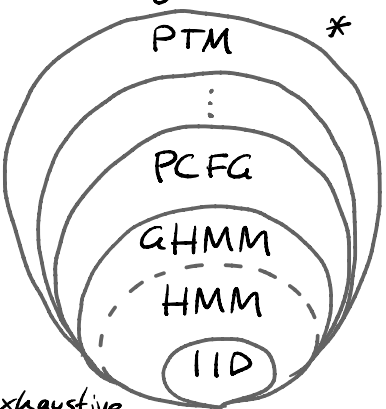
* For what classes of data generating processes do we understand optimal prediction?
* For what scale of neural networks have we identified the structures of optimal prediction?

Processes:

Fix a (computable) stochastic process and consider:

What resources are required to generate this process?

This naturally leads to a hierarchy:



* not exhaustive

Probabilistic Turing Machine

Probabilistic Context-Free Grammar

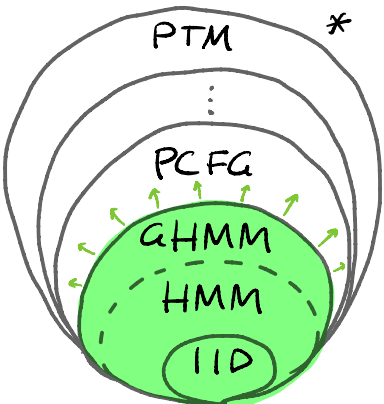
Generalised Hidden Markov Model

Hidden Markov Model

Indep. & Identically Dist.

Processes:

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



* not exhaustive

Probabilistic Turing Machine

Probabilistic Context-Free Grammar

Generalised Hidden Markov Model

Hidden Markov Model

Indep. & Identically Dist.

Neural Networks:

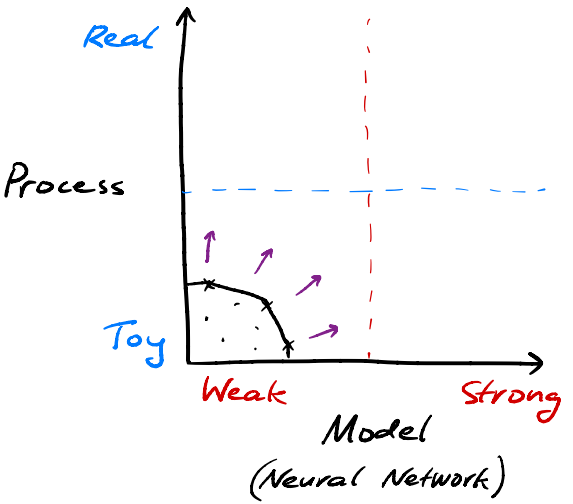
For what scale of neural networks have we identified the structures of optimal prediction?

* Transformer: ~ 10^5 parameters
* LSTM: ~ 10^5 parameters
* RNN: ~ 10^4 parameters
* GRU: ~ 10^5 parameters

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

Overview – where is the program at?

We are working to push the Pareto frontier:



Overview – where is the program at?

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

Weak Strong

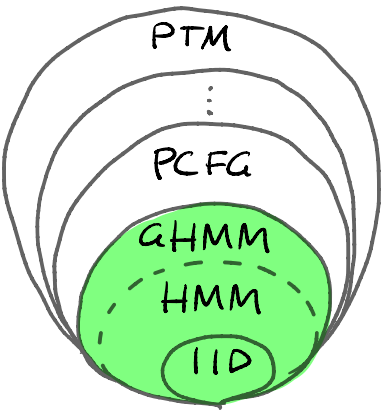
Model

(Neural Network)

Scope – what will we tackle today?

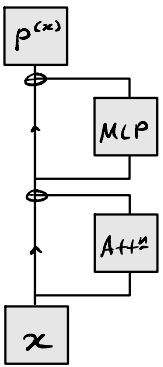
We will take a slightly narrower view and consider:

Processes



GHMMs

Models



Transformers

Plan:

1) Foundations – Hidden Markov Models [lecture/exercises]

2) Foundations – Prediction [lecture/exercises]

3) Applications – Designing Processes [tutorial]

4) Applications – Belief geometry in transformers [reading]

5) Applications – How do transformers do it? [tutorial]

6) Q & A with Adam Shai and/or Paul Riechers

Questions?