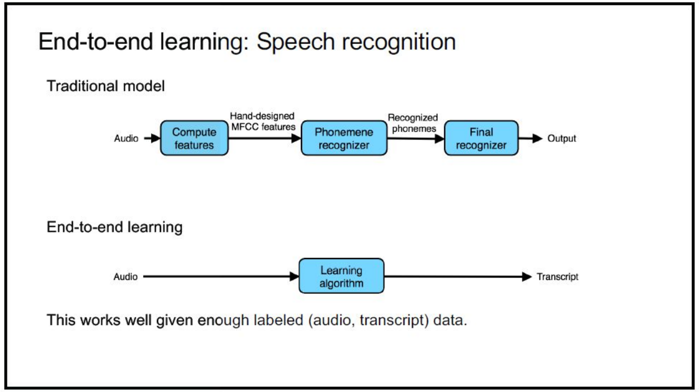
Should attention be all we need?

The epistemic and ethical implications of unification in machine learning

Nic Fishman and Leif Hancox-Li
@njwfish @struthious

What we mean by unification

- Two waves of unification:
  - End-to-end learning
  - Transformers
- Result: similar model architectures used across vastly different domains and tasks

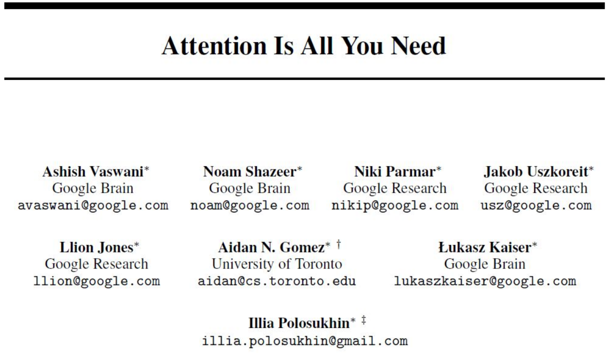


End-to-end learning: Speech recognition

Traditional model

End-to-end learning

This works well given enough labeled (audio, transcript) data.



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

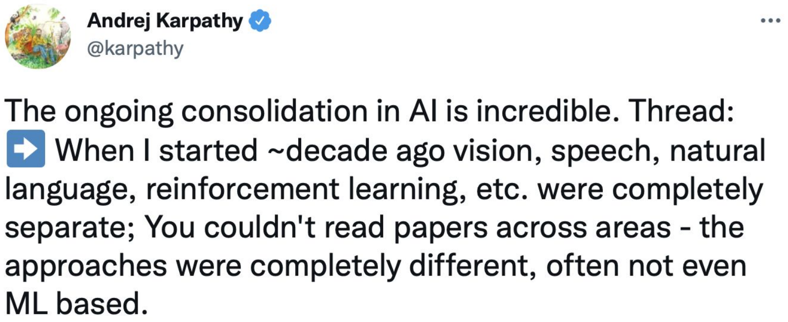
Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

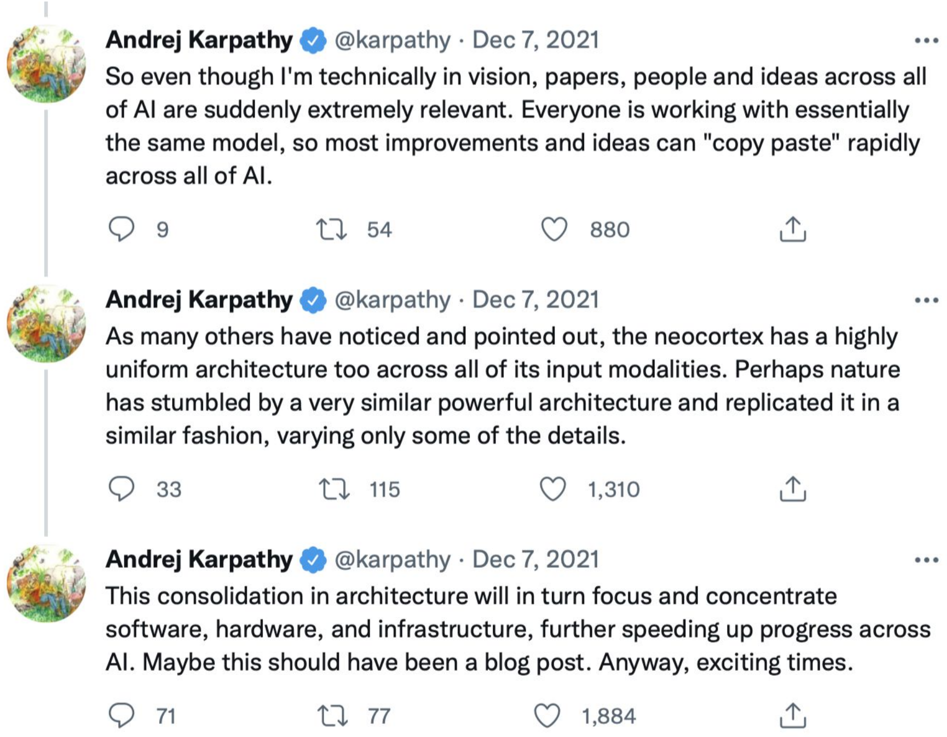
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Pro-unification views in ML



Andrej Karpathy @karpathy

The ongoing consolidation in AI is incredible. Thread: ➡️ When I started ~decade ago vision, speech, natural language, reinforcement learning, etc. were completely separate; You couldn't read papers across areas - the approaches were completely different, often not even ML based.



Andrej Karpathy @karpathy · Dec 7, 2021

So even though I'm technically in vision, papers, people and ideas across all of AI are suddenly extremely relevant. Everyone is working with essentially the same model, so most improvements and ideas can "copy paste" rapidly across all of AI.

9 54 880

Andrej Karpathy @karpathy · Dec 7, 2021

As many others have noticed and pointed out, the neocortex has a highly uniform architecture too across all of its input modalities. Perhaps nature has stumbled by a very similar powerful architecture and replicated it in a similar fashion, varying only some of the details.

33 115 1,310

Andrej Karpathy @karpathy · Dec 7, 2021

This consolidation in architecture will in turn focus and concentrate software, hardware, and infrastructure, further speeding up progress across AI. Maybe this should have been a blog post. Anyway, exciting times.

71 77 1,884

Our analysis of unification

- What benefits are proponents of unification seeing in it? Are they real?
- What risks (epistemic or ethical) does unification bring?

Possible Epistemic Benefits

Ontological Unification: Are neural networks like brains?

Ontological unification in the sciences: Finding a common material/mechanical/structural basis that underlies diverse phenomenon.

- Fundamental particles/forces in physics
- Darwinism

Ontological unification in ML: Finding a common structural basis that underlies both human and artificial intelligences

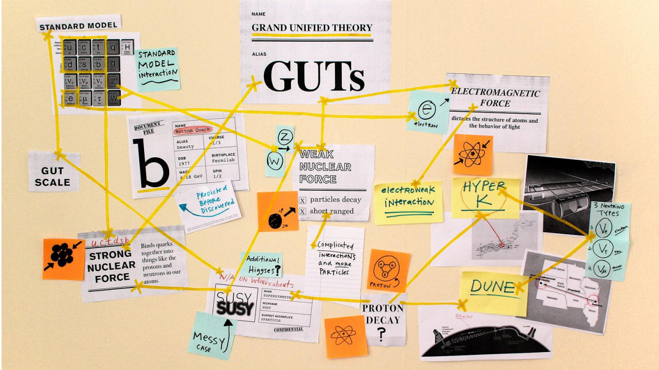
Problems:

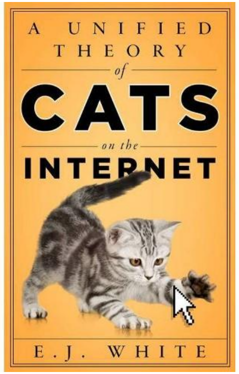
- Current waves of unification based on biologically implausible neural network architectures
- In history of ML, researchers motivated mostly by ‘pragmatic’ concerns, not biological plausibility

Explanatory Unification: Does unification help us explain aspects of neural networks?

Explanatory unification is often a desirable result of unification in science.

- Explain more phenomena with fewer models/theories
- Does unification in ML lead to explanatory unification?
    - 3 possible types of explanatory unification in ML





Explanatory Unification$_1$: Does unification help explain similarities between AI and human intelligence?

What needs explaining: Similarities between current AI systems and human intelligence.

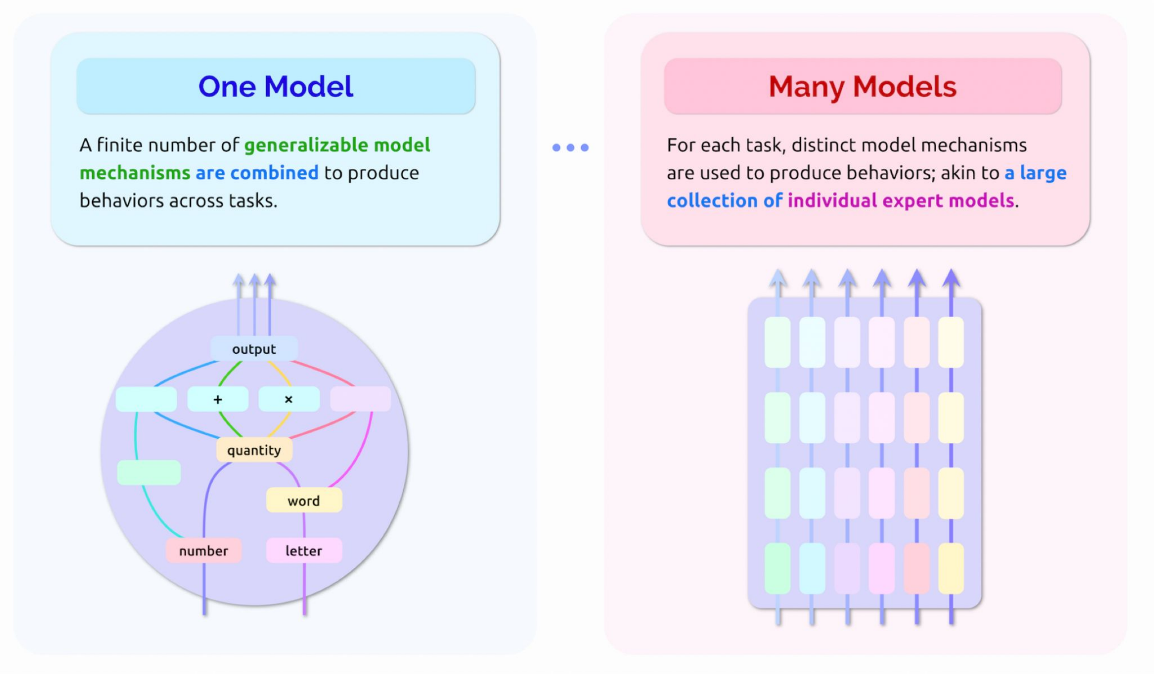
Potential explanation: Because they both share [X ontology].

But: Potential explanation works only if brains and unified AI models are sufficiently ontologically similar.

Explanatory Unification$_2$: Does unification help explain through shared “inductive biases”?

- Artificial neural networks have ‘inductive biases’ that underlie their accuracy.
- To date: there is no unified explanation of what these biases are and why they allow neural networks to learn and generalize (these questions are at the heart of the theory of deep learning).
- Unified architectures holds promise if these biases can be understood, as then they could form the basis of a unified explanation.

Explanatory Unification$_3$: Does unification help explain the performance of one system across multiple domains/tasks?



One Model

A finite number of **generalizable model mechanisms are combined** to produce behaviors across tasks.

Many Models

For each task, distinct model mechanisms are used to produce behaviors; akin to **a large collection of individual expert models**.

Where do multi-domain transformers fall on this spectrum?

Unification as Parsimony: Are unified neural networks simpler?

Argument for unity in science: parsimonious theories are more likely to be true

Does this argument transfer to ML?

Reasons to think not:

1. Current wave of unification does not at present look like it’ll lead to ‘simpler’ models. Nor are multimodal models necessarily ‘simpler’ in their mechanisms.
2. ML models are generally not truth-oriented.

Unity of Tools

- Transformers can be thought of as ‘boundary objects’ or ‘trading zones’ between different subfields of ML
- Shared language/shared tools → quicker sharing of ideas, cross-fertilization, etc.

Epistemic Risks

Ignoring Domain Experts

- More ‘general’ models → less need to involve domain experts in model development process
- This poses epistemic risks:
  - Many prominent ML failures are failures to sufficiently consult domain experts
  - “Portability Trap”
- Knock-on ethical risks (more on this later)

Path Dependency

High switching costs for ML infrastructure

Arbitrariness of the success of ML paradigms

} Path Dependence

Methodological Diversity

The epistemic benefits of machine learning are at least partially defined by its epistemic spillovers: the epistemic benefits ML methods offer when deployed in other sciences, for example. Unification poses two risks in this.

- **Methodological triangulation**: multiple methods giving the same answer forms a more robust evidentiary basis.
- **The effectiveness of diverse problem solvers**: different methods will be differentially good in different settings, so having a diverse array of capable methods is useful.

Increased Black-Boxing

- Less feature engineering → harder to understand
- Mechanisms of larger, unified, multi-modal models are often more opaque than in simpler, more specialized models.
- Also an ethical risk

Ethical Risks and Benefits

Ethical Benefits

- Ethical benefits of unification in machine learning seem most likely to come from **open-sourcing** ML models which are understandable by a wide variety of people (because the models are common tools), as Hugging Face does.
- But the **“many eyeballs” model of security is imperfect**, and can open up systemic risks, as in the log4j vulnerability most recently.

Ethical Risks

- Ethical implications of epistemic risks
- Easier to ignore marginalized perspectives (and ignore harms)
- Further centralization of power
- The “algorithmic leviathan”
- Epistemic homogeneity and social welfare

Conclusion

What’s good/bad about unification?

- We identified a couple of trends towards unification in ML, and ask whether they are good
- We analysed the possible epistemic risks/benefits from a philosophy of science perspective
- We also identified potential ethical risks/benefits