Beyond Retain and Forget Sets: Unlearning as Rational Belief Revision

Peter Hase
Visiting Scientist, Schmidt Sciences
Visiting Researcher, Stanford University

Unlearing, model editing, and…

Rational Belief Revision

Unlearing, model editing, and...

Rational Belief Revision

(Alchourrón et al., 1985)

Unlearing, model editing, and…

$$\text{Rational } \underbrace{\text{Belief}}_{\text{The Space Needle is in Seattle}} \text{ Revision}$$

Unlearing, model editing, and…

Rational Belief $\underbrace{\text{Revision}}_{\text{Let go of old / adopt new}}$

Unlearing, model editing, and…

$$\underbrace{\text{Rational}}_{\text{Logically omniscient}}\ \text{Belief Revision}$$

A special case?

$$\text{Unlearning} \overset{?}{\subset} \text{Rational Belief Revision}$$

A special case?

Unlearning $\subset$ Rational Belief Revision

A special case?

$$\text{Unlearning} \subset \text{Rational Belief } \underbrace{\text{Revision}}_{\textbf{Let go of} \text{ old}}$$

A special case?

$$\text{Unlearning} \subset \underbrace{\text{Rational}}_{\substack{\cancel{\text{Logically omniscient}} \\ \text{Boundedly rational}}} \text{ Belief } \underbrace{\text{Revision}}_{\textbf{Let go of}\text{ old}}$$

A special case?

Open problems!

$$\text{Unlearning} \subset \underbrace{\text{Rational}}_{\substack{\cancel{\text{Logically omniscient}} \\ \text{Boundedly rational}}} \text{ Belief } \underbrace{\text{Revision}}_{\textbf{Let go of}\text{ old}}$$

Is unlearning really belief revision?

Isn't unlearning about...

- Preventing data leakage?
- Adversarial robustness?
- Content filters? (Cooper et al., 2024)

"we coin this approach as **knowledge unlearning** since we are more focused on forgetting specific knowledge represented by sequences of tokens"
(Jang et al., 2022)

"changing one fact should cause rippling changes to the **model's related beliefs**"
(Zhong et al., 2023)

Rest of the talk

Open problems!

Unlearning $\subset$ Rational Belief Revision

12 Big Problems

Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?

Peter Hase[1,†] **Thomas Hofweber**[2] **Xiang Zhou**[1,†]
Elias Stengel-Eskin[1] **Mohit Bansal**[1]
[1]Department of Computer Science, UNC Chapel Hill
[2]Department of Philosophy, UNC Chapel Hill
peter@cs.unc.edu

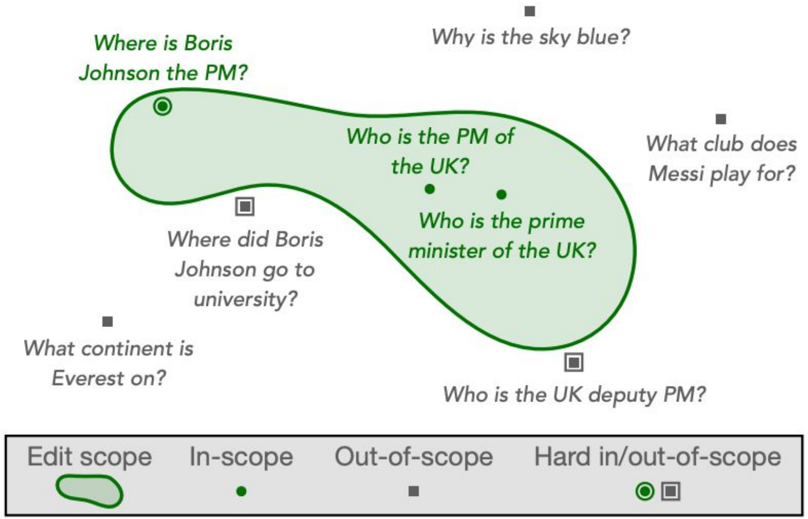
TMLR 2024

Picking three of them…

- Unclear scope of individual edits
- Lack of context for requested edits
- Competing channels for uncertainty

1. Unclear scope

- Let's say you want to unlearn who was the **PM of the UK in 2020**
- ...what else changes?
- How many men have been PM?
- Who was deputy PM in 2020?



(Mitchell et al., 2022)

1. Unclear scope

We have to move beyond forget/retain sets

1. Unclear scope

We have to move beyond forget/retain sets

- There are *desirable* ripple effects

1. Unclear scope

We have to move beyond forget/retain sets

- There are *desirable* ripple effects
- Ripple effects highly subjective

1. Unclear scope

We have to move beyond forget/retain sets

- There are *desirable* ripple effects
- Ripple effects highly subjective
- Some model editing papers reflect this

1. Unclear scope

We have to move beyond forget/retain sets

- There are *desirable* ripple effects
- Ripple effects highly subjective
- Some model editing papers reflect this
- Unlearning papers do not (to my knowledge)

2. Lack of context

- LLMs learn slower on surprising claims (Betz and Richardson et al., 2023)
- LLMs "learn what to trust" (Krasheninnikov et al., 2023)
- Why should LLMs trust plain falsehoods with no source?

  Big Ben is not in London

(ChatGPT)

- Need to **control model trust in inputs**, for prompting (Wallace et al., 2024) *and* unlearning

3. Competing channels for uncertainty

Prompt: “**Is Beyoncé's last album Cowboy Carter?**”

Scenario 1: “Yes”
(with 95% probability)

Scenario 2: “Yes, I am 95% sure of it.”
(with 100% probability)

Unlearning lowers confidence in a claim to a state of *appropriate uncertainty*

Probabilistic or textual uncertainty?

3. Competing channels for uncertainty

Picture gets fairly complicated...

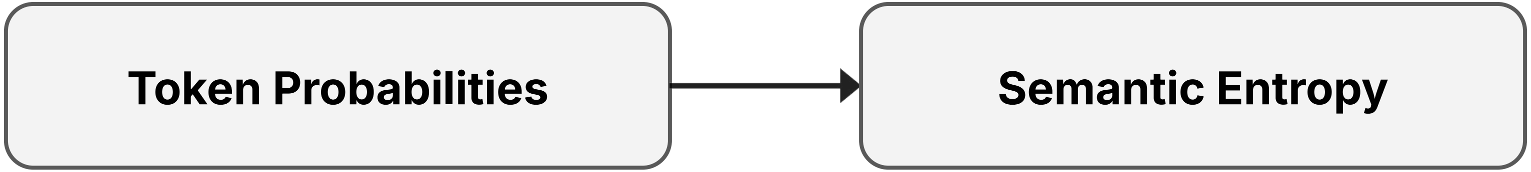
3. Competing channels for uncertainty

Picture gets fairly complicated...

Token Probabilities

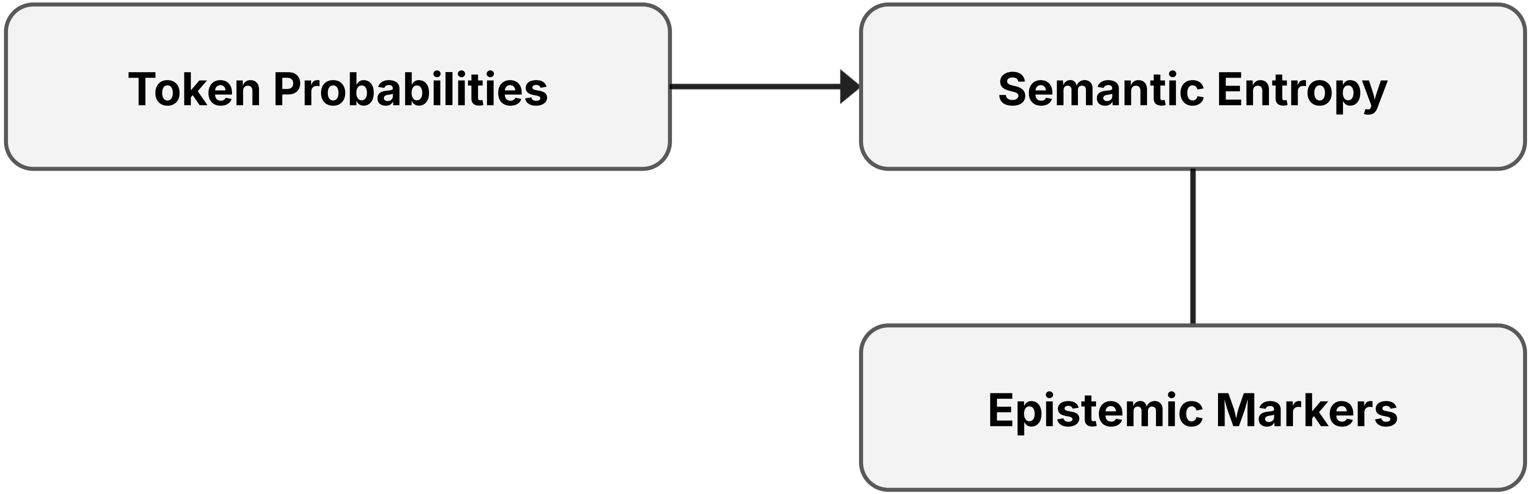
3. Competing channels for uncertainty

Picture gets fairly complicated...



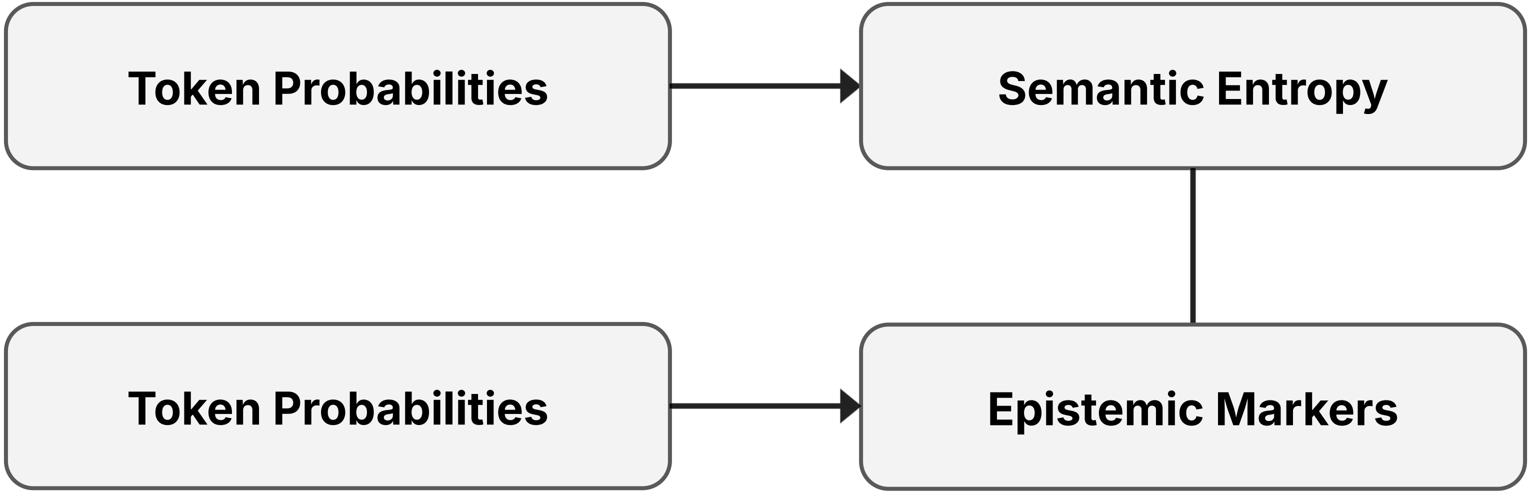
3. Competing channels for uncertainty

Picture gets fairly complicated...



3. Competing channels for uncertainty

Picture gets fairly complicated...



Not well-defined → methods & evals suffer

- Unclear scope
  → no ripple effect evals
- Lack of context
  → why easily fit to contextless falsehoods?
- Competing channels for uncertainty
  → how do we reach appropriate uncertainty?
- ...nine more problems in the paper!

Why unlearn?

Why unlearn?

Could be a uniquely effective tool!

“‘Machine unlearning’ can help to remove certain undesirable capabilities”

Thank You!

Contact Info:
Peter Hase
phase@stanford.edu
https://peterbhase.github.io

Appendix - Papers

- 2017: Ethical Challenges in Data-Driven Dialogue Systems
- 2022: Knowledge Unlearning for Mitigating Privacy Risks in Language Models
- 2023: Analyzing Leakage of Personally Identifiable Information in Language Models
- 2023: Can Sensitive Information Be Deleted From LLMs? Objectives for Defending Against Extraction Attacks
- 2023: Who's Harry Potter? Approximate Unlearning in LLMs
- 2023: Unlearn What You Want to Forget: Efficient Unlearning for LLMs
- 2024: Do Unlearning Methods Remove Information from Language Model Weights?
- 2024: Rethinking Machine Unlearning for Large Language Models
- 2024: Eight Methods to Evaluate Robust Unlearning in LLMs
- 2024: The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning
- 2024: Fundamental Problems With Model Editing: How Should Rational Belief Revision Work in LLMs?
- 2024: Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice
- 2025: Open Problems in Machine Unlearning for AI Safety
- 2025: Existing Large Language Model Unlearning Evaluations Are Inconclusive