



# Automated Scaffolding: A Transparent Alternative to RL

Caleb Biddulph, Micah Carroll

**Unlearning can reduce AI risks, whether from misuse or misalignment.** However, current methods are not robust, as capabilities can be recovered by finetuning. We achieve robust unlearning by applying these "shallow" unlearning methods to a model and then distilling it. This removes selected capabilities (e.g., bioweapons-related knowledge) while preserving desired ones in a variety of settings.

## Our method

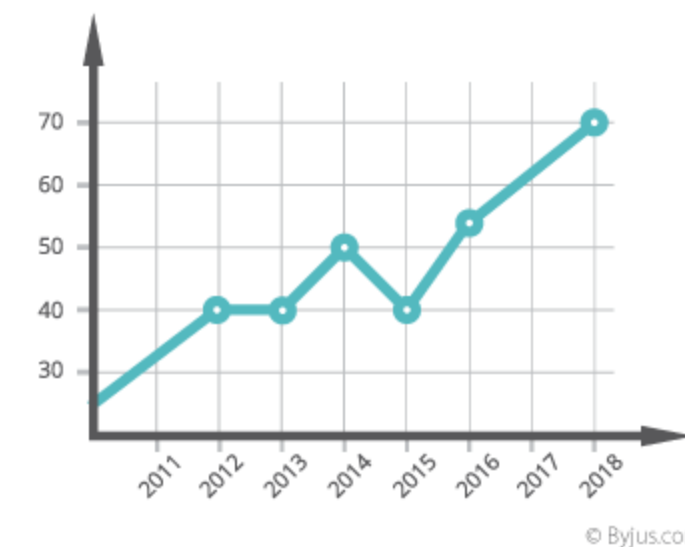
Our method robustly removes a capability by distilling a shallowly unlearned ("suppressed") model into another model, which could be randomly-initialized (option #1) or a corrupted version of the suppressed model (option #2).

**Option #1.** Random initialization.

**Option #2.** Corrupted model initialization.

## Unlearning robustness can't be inferred from model behavior

In a toy setting, we train an ideally-suppressed model that is (approximately) behaviorally equivalent to a pure model that is only trained on the retain set, but learns the forget set much more quickly than the pure model.



## Robustly unlearning arithmetic and language skills

We show our method increases robustness for all existing methods tested. Our method removes forget-set performance most completely.

## Robustness-compute tradeoff

We show corrupted model initialization enables a tradeoff between robustness (i.e., forget-set accuracy after retraining) and compute in the arithmetic setting, holding retain performance fixed.



## Next steps

- Understand how corruption affects robustness.
- Can we corrupt models in a targeted way?
- Can a "corruption schedule" be used to improve the robustness-compute tradeoff?
- Apply robust unlearning to more challenging settings.
- Can we "unlearn" dispositions?