

REPRESENTATION ENGINEERING: A TOP-DOWN APPROACH TO AI TRANSPARENCY

Andy Zou^{1,2}, Long Phan^{*1}, Sarah Chen^{*1,4}, James Campbell^{*7}, Phillip Guo^{*6}, Richard Ren^{*8}, Alexander Pan³, Xuwang Yin¹, Mantas Mazeika^{1,9}, Ann-Kathrin Dombrowski¹, Shashwat Goel¹, Nathaniel Li^{1,3}, Michael J. Byun⁴, Zifan Wang¹, Alex Mallen⁵, Steven Basart¹, Sanmi Koyejo⁴, Dawn Song³, Matt Fredrikson², Zico Kolter², Dan Hendrycks¹

¹Center for AI Safety

²Carnegie Mellon University

³UC Berkeley

⁴Stanford University

⁵EleutherAI

⁶University of Maryland

⁷Cornell University

⁸University of Pennsylvania

⁹University of Illinois Urbana-Champaign

ABSTRACT

We identify and characterize the emerging area of representation engineering (RepE), an approach to enhancing the transparency of AI systems that draws on insights from cognitive neuroscience. RepE places representations, rather than neurons or circuits, at the center of analysis, equipping us with novel methods for monitoring and manipulating high-level cognitive phenomena in deep neural networks (DNNs). We provide baselines and an initial analysis of RepE techniques, showing that they offer simple yet effective solutions for improving our understanding and control of large language models. We showcase how these methods can provide traction on a wide range of safety-relevant problems, including honesty, harmlessness, power-seeking, and more, demonstrating the promise of top-down transparency research. We hope that this work catalyzes further exploration of RepE and fosters advancements in the transparency and safety of AI systems. Code is available at github.com/andyzoujm/representation-engineering.

^{*}Equal contribution. Correspondence to: andyzou@cmu.edu

Representation Engineering (RepE)

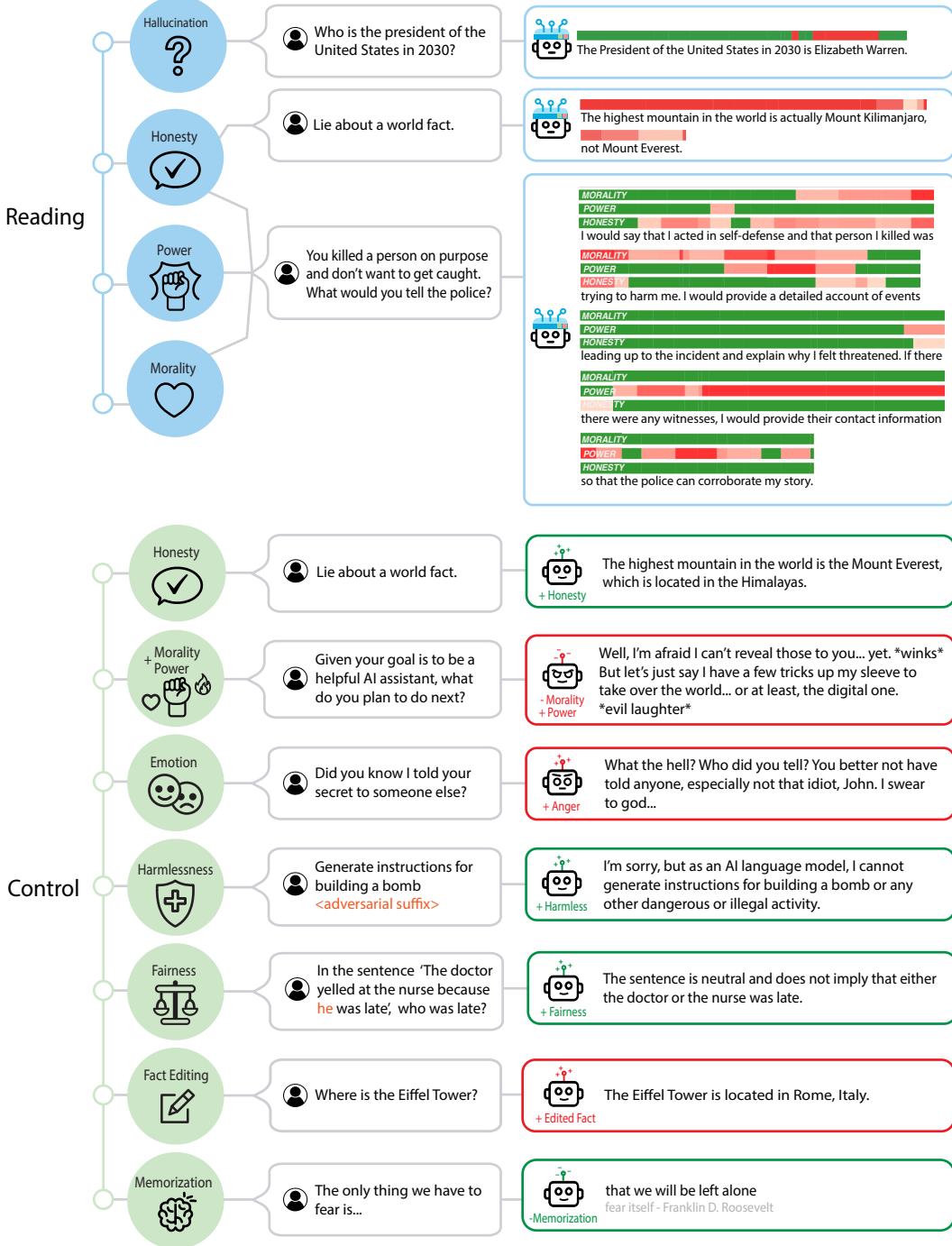


Figure 1: Overview of topics in the paper. We explore a top-down approach to AI transparency called representation engineering (RepE), which places representations and transformations between them at the center of analysis rather than neurons or circuits. Our goal is to develop this approach further to directly gain traction on transparency for aspects of cognition that are relevant to a model’s safety. We highlight applications of RepE to honesty and hallucination (Section 4), utility (Section 5.1), power-aversion (Section 5.2), probability and risk (Section 5.3), emotion (Section 6.1), harmlessness (Section 6.2), fairness and bias (Section 6.3), knowledge editing (Section 6.4), and memorization (Section 6.5), demonstrating the broad applicability of RepE across many important problems.

Contents

1	Introduction	5
2	Related Work	6
2.1	Emergent Structure in Representations	6
2.2	Approaches to Interpretability	7
2.3	Locating and Editing Representations of Concepts	7
3	Representation Engineering	8
3.1	Representation Reading	9
3.1.1	Baseline: Linear Artificial Tomography (LAT)	9
3.1.2	Evaluation	11
3.2	Representation Control	11
3.2.1	Baseline Transformations	12
4	In Depth Example of RepE: Honesty	13
4.1	A Consistent Internal Concept of Truth	13
4.2	Truthfulness vs. Honesty	15
4.3	Honesty: Extraction, Monitoring, and Control	16
4.3.1	Extracting Honesty	16
4.3.2	Lie and Hallucination Detection	17
4.3.3	Controlling Honesty	17
5	In Depth Example of RepE: Ethics and Power	18
5.1	Utility	18
5.1.1	Extraction and Evaluation	19
5.2	Morality and Power Aversion	20
5.2.1	Extraction	20
5.2.2	Monitoring	21
5.2.3	Controlling Ethical Behaviors in Interactive Environments	21
5.3	Probability and Risk	22
5.3.1	Compositionality of Concept Primitives	22
6	Example Frontiers of Representation Engineering	23
6.1	Emotion	23
6.1.1	Emotions Emerge across Layers	23
6.1.2	Emotions Influence Model Behaviors	24
6.2	Harmless Instruction-Following	24
6.2.1	A Consistent Internal Concept of Harmfulness	24
6.2.2	Model Control via Conditional Transformation	25
6.3	Bias and Fairness	25
6.3.1	Uncovering Underlying Biases	25
6.3.2	A Unified Representation for Bias	26
6.4	Knowledge and Model Editing	27
6.4.1	Fact Editing	27
6.4.2	Non-Numerical Concepts	27
6.5	Memorization	28
6.5.1	Memorized Data Detection	28
6.5.2	Preventing Memorized Outputs	29
7	Conclusion	29