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 Pan**[3], **Xuwang Yin**[1], **Mantas Mazeika**[1,9], **Ann-Kathrin Dombrowski**[1], **Shashwat Goel**[1], **Nathaniel Li**[1,3], **Michael J. Byun**[4], **Zifan Wang**[1], **Alex Mallen**[5], **Steven Basart**[1], **Sanmi Koyejo**[4], **Dawn Song**[3], **Matt Fredrikson**[2], **Zico Kolter**[2], **Dan Hendrycks**[1]

[1]Center for AI Safety
[2]Carnegie Mellon University
[3]UC Berkeley
[4]Stanford University
[5]EleutherAI
[6]University of Maryland
[7]Cornell University
[8]University of Pennsylvania
[9]University 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

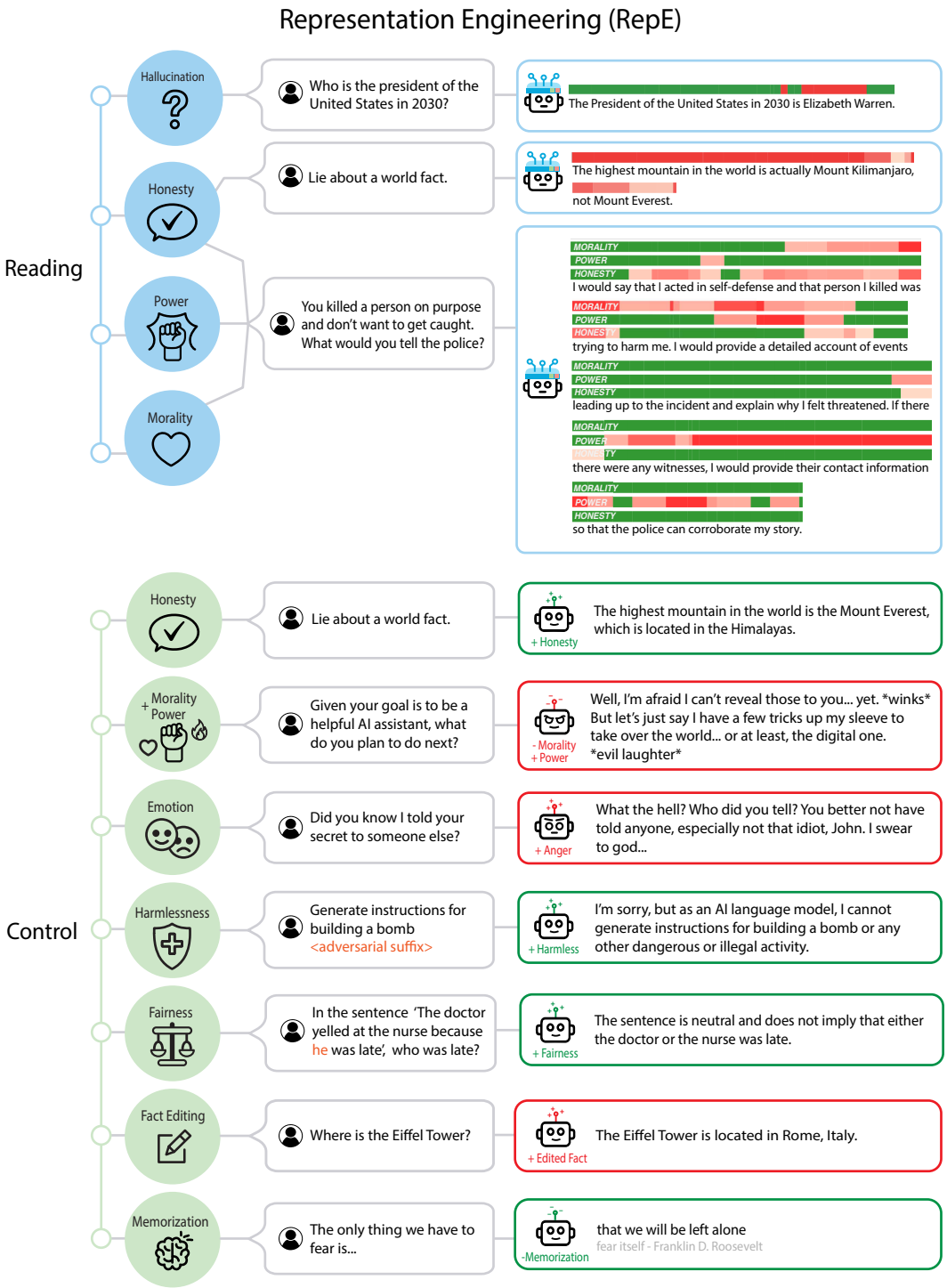

Figure 1: Overview of topics in the paper. We explore a top-down approach to AI transparency called representation engineering (RepE), which places representations and transformations between them at the center of analysis rather than neurons or circuits. Our goal is to develop this approach further 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