Strategic Picture - The strategic picture articulated in this paper – What does superposition mean for interpretability and safety? What would a suitable solution be? How might one solve it? – developed in extensive conversations between authors, and in particular Chris Olah, Tristan Hume, Nelson Elhage, Dario Amodei, Jared Kaplan. Nelson Elhage recognized the potential importance of "enumerative safety", further articulated by Dario. Tristan brainstormed extensively about ways one might solve superposition and pushed Chris on this topic.

Writing - The paper was primarily drafted by Chris Olah, with some sections by Nelson Elhage, Tristan Hume, Martin Wattenberg, and Catherine Olsson. All authors contributed to editing, with particularly significant contributions from Zac Hatfield Dodds, Robert Lasenby, Kipply Chen, and Roger Grosse.

Illustration - The paper was primarily illustrated by Chris Olah, with help from Tristan Hume, Nelson Elhage, and Catherine Olsson.

Citation Information

Please cite as:

```
Elhage, et al., "Toy Models of Superposition", Transformer Circuits Thread, 2022.
```

BibTeX Citation:

```
@article{elhage2022superposition,
   title={Toy Models of Superposition},
   author={Elhage, Nelson and Hume, Tristan and Olsson, Catherine and Schiefer, Nicholas and Henighan,
Tom and Kravec, Shauna and Hatfield-Dodds, Zac and Lasenby, Robert and Drain, Dawn and Chen, Carol and
Grosse, Roger and McCandlish, Sam and Kaplan, Jared and Amodei, Dario and Wattenberg, Martin and Olah,
Christopher},
   year={2022},
   journal={Transformer Circuits Thread},
   note={https://transformer-circuits.pub/2022/toy_model/index.html}
}
```