## DTL Past Working Group Readings

### Winter 2025: Interpretability and/as Computation

* DeepSeek, DeepSeek-V3 Technical Report; Jason Wei et al, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”
* Stephen Wolfram, “How to Think Computationally”; Charles Petzold, *The Annotated Turing*
* Peter Gardenförs, “Dimensions” and “Computational Aspects” in *Conceptual Spaces*; Christopher Gauker, “Regions of Similarity Space”
* Imogen Forbes-Macphail, “The Four-Color Theorem and the Aesthetics of Computational Proof”

### Fall 2024: Interpretability

* Anthropic, “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet”; John Haugeland, “Understanding Natural Language”
* Adrian Mackenzie, “Vectorization and Its Consequences” in *Machine Learners*; Bernhard Rieder, “From Frequencies to Vectors” in *Engines of Order*
* Minyoung Huh et al, “The Platonic Representation Hypothesis”; Jiaang Li et al, “Do Vision and Language Models Share Concepts?”; Jack Merullo et al, “Linearly Mapping from Image to Text Space”; Albert Jiang et al, “Mixtral of Experts”
* Franco Moretti, “Operationalization”

### Spring 2024: Form

* John Haugeland, “Good Old-Fashioned AI” in *Artificial Intelligence: The Very Idea*
* Jonathan Kramnick and Anahid Nersessian, “Form and Explanation”; Matthew Kirschenbaum, “Spec Acts: Reading Form in Recurrent Neural Networks” Ernest
* Nagel and James Newman, *Gödel’s Proof*
* Jacques Derrida, “Law of Genre”; Ralph Cohen, “History and Genre”
* Erwin Panofsky, *Perspective as Symbolic Form*
* Terry Winograd, “Computation, Thought, and Language” in *Understanding Computers and Cognition*
* Leo Breiman, “Statistical Modeling: The Two Cultures”