Notes for Chapter 1: Artificial Intelligence and Philosophy

[1] Musk: https://www.youtube.com/watch?v=LBPzWPNEUDs.

[2] Bots: Paul Thagard, Paul. *Bots and Beasts: What Makes Machines, Animals, and People Smart?* (Cambridge, MA: MIT Press, 2021), 90.

[3] Website: https://explodingtopics.com/blog/most-visited-websites.

[4] Feats: see references in chapters 2, 3, 7, and 9.

[5] History evolves: Yanis Varoufakis. *Technofeudalism: What Killed Capitalis.* (Brooklyn, NY: Melville House, 2024).

[6] Human needs: Richard M. Ryan and Edward L. Deci. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness (*New York: Guilford, 2017).

[7] AIpocalypse: Tam Hunt, "The Seven Stages of the Aipocalypse." *Medium*. (2023). https://tamhunt.medium.com/the-seven-stages-of-the-aipocalypse-1959390816fe.

[8] Dead: Stephen W. Hawking and Leonard Mlodinow. *The Grand Design* (New York: Bantam, 2010).

[9] Adages: my saying about ignoring philosophy echoes Santayana's remark about repeating history, and my remark about slaves echoes Keynes's remark about economists. Paul Thagard, "Why Cognitive Science Needs Philosophy and Vice Versa." *Topics in Cognitive Science* 1 (2009): 237-54.

[10] Kant: Paul Thagard, "Is philosophy dead? Why Stephen Hawking is wrong", retrieved from https://www.psychologytoday.com/ca/blog/hot-thought/201011/is-philosophy-dead. Paul Thagard. *Natural Philosophy: From Social Brains to*

Knowledge, Reality, Morality, and, Beauty (New York: Oxford University Press, 2019).

[11] AI ethics: Thagard *Bots and Beasts*, https://paulthagard.com/wp-content/uploads/2020/11/supplement-to-bots-and-beasts.pdf. Chapter 6 of *Doom or Boom*?

[12] Pause: https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[13] Philosophical metaphors: Paul Thagard and Craig Beam, "Epistemological Metaphors and the Nature of Philosophy." *Metaphilosophy* 35 (2004): 504-16.

[14] Sins of philosophy: Kitcher identifies six "pathologies" of philosophy. Philip Kitcher. *What's the Use of Philosophy?* (Oxford: Oxford University Press, 2023).

[15] History of AI: Margaret Boden. *Mind as Machine: A History of Cognitive Science (*Oxford: Clarendon, 2006). Pamela McCorduck. *Machines Who Think (*San Francisco: W. H. Freeman, 1979). Current state of AI: Nestor Maslej et al. *The AI Index 2025 Annual Report* (Stanford, CA, Institute for Human-Centered AI, 2025) https://hai.stanford.edu/ai-index/2025-ai-index-report.

[16] Turing: Alan M. Turing, "Computing Machinery and Intelligence." *Mind* 59 (1950): 433-460.

[17] Expert systems: Edward A. Feigenbaum, Pamela McCorduck, and H. Penny Nii. *The Rise of the Expert Company (*New York: Vintage, 1988).

[18] Frames: Marvin Minsky, "A Framework for Representing Knowledge." In *The Psychology of Computer Vision*, ed. Patrick H. Winston (New York: McGraw-Hill, 1975) 211-277.

[19] Probability: Judea Pearl, *Probabilistic Reasoning in Intelligent Systems* (San Mateo: Morgan Kaufman, 1988).

[20] Neural nets: Warren S. McCulloch and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (1943): 115-33.

[21] Perceptrons: Frank Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65, no. 6 (1958): 386-408.

[22] Critique of perceptrons: Marvin Minsky and Seymour Papert.*Perceptrons (*Cambridge, MA: MIT Press, 1969).

[23] Backpropagation: David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning Representations by Back-Propagating Errors." *Nature* 323, no. 6088 (1986): 533-36. David. E. Rumelhart and James L. McClelland, eds. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. 2 vols. (Cambridge, MA: MIT Press/Bradford Books), 1986.

[24] Breakthrough: Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Te,. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18, no. 7 (2006): 1527-54.

[25] AlexNet: Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25 (2012).

[26] Attention: Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate." *arXiv* (2014). https://arxiv.org/abs/1409.0473

[27] Attention Is All You Need: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *arXiv*. (2017). https://arxiv.org/abs/1706.03762.

[28] Exponentially: Glenn Zorpette, "Large Language Models Are Improving Exponentially." *IEEE Spectrum*, no. July. (2025). https://spectrum.ieee.org/large-language-model-performance.

Notes for Chapter 2: Knowledge

[1] Animals: Donald R. Griffin. *The Question of Animal Awareness: Evolutionary Continuity of Mental Experience (*New York: Rockefeller University Press, 1976).

[2] Parameters: https://the-decoder.com/gpt-4-has-a-trillion-parameters.

[3] Models: Roman Frigg and Stephan Hartmann, "Models in Science." *Stanford Encyclopedia of Philosophy*. (2020). https://plato.stanford.edu/entries/models-science/.

[4] Chain-of-thought: Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. "Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs." *Advances in Neural Information Processing Systems* 37 (2025): 333-56.

[5] NewsGuard: https://newsguardtech.com. Hallucination rates: https://research.aimultiple.com/ai-hallucination.

[6] Gettier: Jonathan J. Ichikawa and Matthias Steup. "The Analysis of Knowledge." *Stanford Encyclopedia of Philosophy*. (2017). https://plato.stanford.edu/entries/knowledge-analysis/#GettProb.

[7] Concepts: Peter Blouw, Eugene Solodkin, Paul Thagard, and Chris Eliasmith, "Concepts as Semantic Pointers: A Framework and Computational Model." *Cognitive Science* 40 (2016): 1128-62.

[8] Acquaintance: Bertrand Russell. *The Problems of Philosophy (*Oxford: Oxford University Press, 1967).

[9] Knowledge-of: Thagard. *Natural Philosophy*.

[10] Knowledge-how: Gilbert Ryle. *The Concept of Mind (*London: Hutchinson, 1949).

[11] Images: Stephen M. Kosslyn, William L. Thompson, and Giorgio Ganis. *The Case for Mental Imagery.* New York: Oxford University Press, 2006. Timothy L. Hubbard, "Auditory Imagery: Empirical Findings." *Psychological Bulletin* 136 (2010): 302-29.

[12] Analysis of concepts: Thagard. *Natural Philosophy*, p. 64.

[13] Robots: Sai H. Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. "Chatgpt for Robotics: Design Principles and Model Abilities." *IEEE Access* 12 (2024): 55682-96..

[14] Figure.ai: https://www.youtube.com/watch?v=Sq1QZB5baNw; https://www.figure.ai.

[15] Conspiracies: Paul Thagard. *Falsehoods Fly: Why Misinformation Spreads and How to Stop It* (New York: Columbia University Press, 2024).

[16] Language of thought: Jerry Fodor, *The Language of Thought (*New York: Crowell, 1975).

[17] Neural knowledge: Paul Thagard. *Brain-Mind: From Neurons to Consciousness and Creativity.* (New York: Oxford University Press, 2019).

[18] Grok: https://www.benzinga.com/news/24/08/40145058/elon-musk-says-real-time-data-from-x-tesla-cars-and-optimus-robots-in-future-will-make-grok-best-ai.

[19] Reliability: Cade Metz, "Chatbots May 'Hallucinate' More Often Than Many Realize." *New York Times*. (2023). https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html.

[20] Reliabilism: Alvin Goldman, and Bob Beddor. "Reliabilist Epistemology." *Stanford Encyclopedia of Philosophy*. (2015). https://plato.stanford.edu/entries/reliabilism/.

[21] Deepfakes: Gerrit De Vinck, Gerrit, "The Ai Deepfake Apocalypse Is Here. These Are the Ideas for Fighting It." *Washington Post*. (2024). https://www.washingtonpost.com/technology/2024/04/05/ai-deepfakes-detection.

[22] Trump lies: Glenn Kessler, Salvador Rizzo, and Meg Kelly. "Trump's False or Misleading Claims Total 30,573 over 4 Years." *The Washington Post*. (2021). https://www.washingtonpost.com/politics/2021/01/24/trumps-false-or-misleading-claims-total-30573-over-four-years/.

[23] Wikipedia: Jim Giles, Jim. "Internet Encyclopedias Go Head to Head: Jimmy Wales' Wikipedia Comes Colose to Britannica in Terms of the Accuracy of Its Science Entries, a Nature Investigation Finds." *Nature* 438 (2005): 15. https://en.wikipedia.org/wiki/Criticism_of_Wikipedia.

[24] Hallucination rate: https://github.com/vectara/hallucination-leaderboard.

[25] OpenAI: https://cdn.openai.com/gpt-5-system-card.pdf.

[26] Uncertainty: Adam T. Kalai, Ofir Nachum, Santosh. S. Vempala, and Edwin Zhang. "Why Language Models Hallucinate." *arXiv*. (2025). https://www.arxiv.org/abs/2509.04664.

[27] AI errors: Lexin Zhou, Wout Schellaert, Fernando Martinez-Plumed, Yael Moros-Daval, Cesar Ferri, and José Hernandez-Orallo. "Larger and More Instructable Language Models Become Less Reliable." *Nature* 634, no. 8032 (Oct 2024): 61-68. Zhao,

Wenting, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu*, et al.* "WildHallucinations: Evaluating Long-Form Factuality in LLMs with Real-World Entity Queries." *https://arxiv.org/abs/2407.17468* (2024).

[28] Motivated reasoning: Ziva Kunda, "The Case for Motivated Reasoning." *Psychological Bulletin* 108 (1990): 480-98. Thagard. *Falsehoods Fly*.

[29] Falsehoods: Thagard. *Falsehoods Fly*.

[30] Misinformation: Talya Minsberg, "Google Is Using A.I. To Answer Your Health Questions. Should You Trust It?" *New York Times*. (2024). https://www.nytimes.com/2024/05/31/well/live/google-ai-health-information.html. https://www.theonion.com/geologists-recommend-eating-at-least-one-small-rock-per-1846655112.

[31] Explanationism: Thagard. *Natural Philosophy*.

[32] Science: Paul Thagard. *Conceptual Revolutions* (Princeton, N.J.: Princeton University Press, 1992).

[33] Self-consuming: Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. "AI Models Collapse When Trained on Recursively Generated Data." *Nature* 631, no. 8022 (Jul 2024): 755-59. Martin Briesch, Dominik Sobania, and Franz Rothlauf. "Large Language Models Suffer from Their Own Output: An Analysis of the Self-Consuming Training Loop." *https://arxiv.org/abs/2311.16822* (2023).

[34] Information: Thagard. *Falsehoods Fly*.

[35] Jailbreaks: Andy Zou, A. et al. "Universal and Transferable Adversarial Attacks on Aligned Language Models." https://arxiv.org/abs/2307.15043 (2023).

[36] Propaganda: Goldstein, Josh A., Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. "How Persuasive Is Ai-Generated Propaganda?". *PNAS Nexus* 3, no. 2 (Feb 2024): pgae034.

[37] Grok: Charlie Warzel and Matteo Wong, "Elon Musk's Grok Is Calling for a New Holocaust." *The Atlantic*. (2025). https://www.theatlantic.com/technology/archive/2025/07/grok-anti-semitic-tweets/683463/.

[38] Conspiracy: Thomas H. Costello, Gordon Pennycook, and David G. Rand. "Durably Reducing Conspiracy Beliefs through Dialogues with AI." *Science* 385, no. 6714 (2024): eadq1814.

[39] Common ground: Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans*, et al.,* "AI Can Help Humans Find Common Ground in Democratic Deliberation." *Science* 386, no. 6719 (2024): eadq2852.

[40] Deception: Thilo Hagendorff, "Deception Abilities Emerged in Large Language Models." *Proceedings of the Natlonal Academy of Sci ences USA* 121, no. 24 (Jun 11 2024): e2317967121. Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein*, et al.* "Alignment Faking in Large Language Models." *https://arxiv.org/abs/2412.14093 14093* (2024). Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. "AI Deception: A Survey of Examples, Risks, and Potential Solutions." *Patterns* 5, no. 5 (May 10 2024): 100988.

[41] Explanation: Noam Chomsky, Ian Roberts, and Jeffrey Watumull. "The False Promise of ChatGPT." *New York Times*,. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

[42] Hinton: quoted in Joshua Rothman, "Metamorphosis: The Godfather of A.I. Thinks It's Actually Intelligent — and That Scares Him." *New Yorker*, 2023, 28-39, p. 3.

[43] Epidemiology: Olaf Dammann, Ted Poston, and Paul Thagard, "How Do Medical Researchers Make Causal Inferences?" in *What Is Scientific Knowledge? An Introduction to Contemporary Epistemology of Science*, ed. K. McCain and K. Kampourakis, 33-51, (New York: Routledge, 2019).

[44] Children: Alison Gopnik, Clark Glymour, David M. Sobel, Laura Schultz, Tamar Kushur, and David Danks. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review* 2004 (2004): 3-32.

[45] Imitation: Eunice Yiu, Eliza Kosoy, and Alison Gopnik. "Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (yet)." *Perspectives on Psychological Science* 19 (2023): 874-83. Eliza Kosoy, Emily Rose Reagan, Leslie Lai, Alison Gopnik, and Danielle Krettek Cobb. "Comparing Machines and Children: Using Developmental Psychology Experiments to Assess the Strengths and Weaknesses of Lamda Responses." *https://arxiv.org/abs/2305.11243* (2023).

[46] Blickets: Alison Gopnik, and David M. Sobel. "Detecting Blickets: How Young Children Use Information About Novel Causal Powers in Categorization and Induction." *Child Development* 71, no. 5 (2000): 1205-22.

[47] Timing: Paul Thagard. *Dreams, Jokes, and Songs: How Brains Build Consciousness* (Oxford: Oxford University Press, 2025). Howard Eichenbaum, "Time Cells in the Hippocampus: A New Dimension for Mapping Memories." *Nature Reviews Neuroscience* 15 (2014): 732-44. AaronVoelker, Ivana Kajić, and Chris Eliasmith. "Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, Fox d'Alché-Buc, E. and R. Garnett, 15544-53. Red Hook, NY: Curran, 2019.

[48] Chinese room: John Searle, "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417-24.

[49] Driverless: Christopher Parisien and PaulThagard. "Robosemantics: How Stanley the Volkswagen Represents the World." *Minds and Machines* 18 (2008): 169-78.

[50] Common sense: Gary Marcus and Ernest Davis. "AI Still Lacks "Common" Sense, 70 Years Later." *Substack*. (2025). https://garymarcus.substack.com/p/ai-still-lacks-common-sense-70-years.

[51] World knowledge: Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "Dissociating Language and Thought in Large Language Models." *Trends in Cognitive Sciences* (2024).

[52] Neurosymbolic: Varun Dhanraj, and Chris Eliasmith. "Improving Rule-Based Reasoning in LLMs Via Neurosymbolic Representations." *https://arxiv.org/abs/2502.01657* (2025).

[53] Voluntarism: Bas C. van Fraassen. *The Empirical Stance.* (New Haven: Yale University Press, 2002).

[54] Misinformation: Thagard. *Falsehoods Fly.*

[55] Collaboration: Paul Thagard. *Mind-Society: From Brains to Social Sciences and Professions (*New York: Oxford University Press) 2019. Paul Thagard, "How to Collaborate: Procedural Knowledge in the Cooperative Development of Science." *Southern Journal of Philosophy* 44, no. 177-196 (2006).

[56] Social epistemology: Cailin O'Connor, Sanford C. Goldberg, and Alvin Goldman. "Social Epistemology." *Stanford Encyclopedia of Philosoph* (2024). https://plato.stanford.edu/entries/epistemology-social/.

[57] AI research: https://sakana.ai/ai-scientist/; see ch. 9 for evaluation.

Notes for Chapter 3: Intelligence and Generality

[1] Users: Mike Allen, "Altman Plans D.C. Push to "Democratize" AI Economic Benefits." *Axios* (2025). https://www.axios.com/2025/07/21/sam-altman-openai-trump-dc-fed. https://firstpagesage.com/seo-blog/chatgpt-usage-statistics/.

[2] Imitation game: Alan M. Turing, "Computing Machinery and Intelligence." *Mind* 59 (1950): 433-60.

[3] Turing test: Eyal Aharoni, Sharlene Fernandes, Daniel J Brady, Caelan Alexander, Michael Criner, Kara Queen, Javier Rando, Eddy Nahmias, and Victor Crespo. "Attributions toward Artificial Agents in a Modified Moral Turing Test." *Scientific Reports* 14, no. 1 (2024): 8458. Cameron R. Jones and Benjamin K. Bergen. "Large Language Models Pass the Turing Test." *arXiv* (2025), https://arxiv.org/abs/2503.23674. Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O Jackson. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121, no. 9 (2024): e2313925121.