BIBLIOGRAPHY

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6:52, 138–52, 160.

Alam et al. (2004). A comparison of experimental designs in the development of a neural network simulation metamodel. *Simulation Modelling Practice and Theory* 12 (7–8).

Alvarez, Maria (2016). Reasons for action: justification, motivation, explanation. *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl>.

Armstrong, D.M. (1997). *A World of States of Affairs*. Cambridge University Press.

Audi, Robert (2001). *The Architecture of Reason: The Structure and Substance of Rationality*. Oxford University Press.

Axelrod, V., Rozier, C., Malkinson, T.S., Lehongre, K., Adam, C., Lambrecq, V., Navarro, V., and Naccache, L. (2019). Face-selective neurons in the vicinity of the human fusiform face area. *Neurology* 92 (4):197–8.

Bard, Nolan, Foerster, Jakob N., Chandar, Sarath, Burch, Neil, Lanctot, Marc, Song, H. Francis, Parisotto, Emilio, Dumoulin, Vincent, Moitra, Subhodeep, Hughes, Edward, Dunning, Iain, Mourad, Shibl, Larochelle, Hugo, Bellemare, Marc G., and Bowling, Michael (2019). The Hanabi Challenge: a new frontier for AI research. Retrieved from: https://arxiv.org/pdf/1902.00506v1.pdf.

Bostrom, Nick (ed.) (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Brandom, Robert B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Brentano, F. (1874 [1911]). *Psychology from an Empirical Standpoint*. London: Routledge and Kegan Paul.

Bringsjord, Selmer and Naveen Sundar Govindarajulu (2018). Artificial Intelligence. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), URL = <https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/>.

Brown, Jessica and Cappelen, Herman (eds.) (2011). *Assertion: New Philosophical Essays*. Oxford University Press.

Buckner, Cameron (2018). Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese* (12):1–34.

Buckner, Cameron (2019). Deep learning: a philosophical introduction. *Philosophy Compass* 14 (10).

Burge, Tyler (1979). Individualism and the mental. *Midwest Studies in Philosophy* 4 (1):73–122.

Cappelen, Herman (2011). Against Assertion. In Jessica Brown and Herman Cappelen (eds.), *Assertion: New Philosophical Essays*. Oxford University Press.

Cappelen, Herman and Dever, Josh (2018). *Puzzles Of Reference*. Oxford: Oxford University Press.

Chalmers, David (2006). The foundations of two-dimensional semantics. In Garcia-Carpintero and Josep Macia (eds.), *Two-Dimensional Semantics: Foundations and Applications*. Oxford University Press. pp. 55–140.

Chiang, Ted. (2002). Stories of Your Life And Others, Tor.

Clark, Andy and Chalmers, David J. (1998). The extended mind. *Analysis* 58 (1):7–19.

Cohnitz, Daniel and Haukioja, Jussi (2013). 'Meta-externalism vs. meta-internalism in the Study of Reference', *Australasian Journal of Philosophy* 91:475–500.

Cole, David (2014). The chinese room argument. *Stanford Encyclopedia of Philosophy,* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2020/entries/chinese-room>.

Crane, Tim (1998). Intentionality as the mark of the mental. In A. O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Dancy, Jonathan (2000). *Practical Reality*. Oxford University Press.

Davidson, Donald (1973). Radical interpretation. *Dialectica* 27 (1):314–28.

Davidson, Donald (1987). Knowing one's own mind. in *Proceedings and Addresses of the American Philosophical Association* 60:441–58.

Donnellan, Keith S. (1966). Reference and definite descriptions. *Philosophical Review* 75 (3):281–304.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. [arXiv:1702.08608v2]. Retrieved from https://arxiv.org/pdf/1702.08608.pdf.

Dretske, Fred (1980). The intentionality of cognitive states. In D. Rosenthal (ed.), *The Nature of Mind,* Oxford: Oxford University Press.

Eliasmith, C., (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Evans, Gareth (1973). The causal theory of names. *Aristotelian Society Supplementary Volume* 47 (1):187–208.

Evans, Gareth (1982). *The Varieties of Reference*. Oxford University Press.

Feldman, Richard. (1998). Principles of charity. In Peter Klein and Richard Foley (eds.), *Routledge Encyclopedia of Philosophy*, available at <https://www.rep.routledge.com/articles/thematic/charity-principle-of/v-1>

Floridi, Luciano (2011). *The Philosophy of Information*. Oxford University Press.

Floridi, Luciano, Cowls, Josh, Beltrametti, Monica, Chatila, Raja, Chazerand, Patrice, Dignum, Virginia, Luetge, Christoph, Madelin, Robert, Pagallo, Ugo, Rossi, Francesca, Schafer, Burkhard, Valcke, Peggy, and Vayena, Effy (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28 (4):689–707.

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

Fodor, Jerry (1990). A theory of content. In *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, Bradford Book.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford, UK: Oxford University Press.

Goldberg, S. (ed.) (forthcoming). *The Oxford Handbook of Assertion*. Oxford University Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. Available at https://www.deeplearningbook.org.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* 27, available at <https://papers.nips.cc/paper/2014>

Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decisionmaking and a 'right to explanation.' Presented at the ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

Gray, A. (2016). Minimal descriptivism. *Review of Philosophy and Psychology, 7*(2), 343–364. doi:10.1007/s13164-014-0202-7

Hanks, Peter (2015). *Propositional Content*. Oxford University Press.

Hawthorne, John and Manley, David (2012). *The Reference Book*. Oxford University Press.

Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating visual explanations. In *European Conference on Computer Vision* (pp. 3–19). Springer.

Huang, Sandy H., Zambelli, Martina, Kay, Jackie, Martins, Murilo F., Tassa, Yuval, Pilarski, Patrick M., and Hadsell, Raia (2019). Learning Gentle Object Manipulation with Curiosity-Driven Deep Reinforcement Learning. Retrieved from: https://arxiv.org/abs/1903.08542.

Kaminski, Margot E. (2019). The right to explanation, explained (June 15, 2018). U of Colorado Law Legal Studies Research Paper No. 18–24; *Berkeley Technology Law Journal* 34 (1): 189-219. Available at SSRN: https://ssrn.com/abstract=3196985 or http://dx.doi.org/10.2139/ssrn.3196985.

King, Jeffrey C. (2007). The Nature and Structure of Content. Oxford University Press.

King, Jeffrey C.; Soames, Scott & Speaks, Jeff (2014). *New Thinking About Propositions*. Oxford University Press.

Kripke, Saul A. (1980). *Naming and Necessity*. Harvard University Press.

Kroon, Frederick W. (1987). Causal descriptivism. *Australasian Journal of Philosophy* 65 (1):1–17.

LeCun, Yann (2017). My take on Ali Rahimi's 'Test of Time' award talk at NIPS. Retrieved from: https://www2.isye.gatech.edu/~tzhao80/Yann_Response.pdf.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553):436–44.

Lewis, David K. (1984). Putnam's paradox. *Australasian Journal of Philosophy* 62 (3):221–36.

Lin, Patrick, Abney, Keith, and Jenkins, Ryan (eds.) (2017). *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press.

Lockwood, Michael (1971). Identity and reference. In Milton Karl Munitz (ed.), *Identity and Individuation*. New York: New York University Press, pp. 199–211.

López-Rubio, Ezequiel (2018). Computational functionalism for the deep learning era. *Minds and Machines* 28 (4):667–88.

Mele, Alfred R. (2003). *Motivation and Agency*. Oxford University Press.

Miller, Tim (2018). Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence* 267:1–38.

Millikan, R.G. (1984). *Language, Thought and Other Biological Objects*. Cambridge, Mass.: MIT Press.

Millikan, R.G. (1989a). In defense of proper functions. *Philosophy of Science* 56 (2):288–302.

Millikan, R.G. (1989b). Biosemantics. *Journal of Philosophy* 86:281–97.

Mueller, Shane T, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein (2019). Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI. DARPA XAI project, available at url=< https://arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf>.

Mulligan, Kevin (2007). Facts. *Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/facts/>.

Neander, Karen (1991). Functions as selected effects. *Philosophy of Science* 58:168–84.

Neander, Karen (1996). Swampman meets swampcow. *Mind and Language* 11 (1):70–130.

Neander, Karen (2006). Content for cognitive science. In G. McDonald and D. Papineau (eds.), *Teleosemantics*, Oxford: Oxford University Press, pp. 167–94.

Nefdt, Ryan M. (2020). A Puzzle concerning Compositionality in Machines. *Minds and Machines* 30 (1):47–75.

Nyholm, Sven and Smids, Jilles (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice* 19 (5):1275–89.

Páez, Andrés (2019). The pragmatic turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* 29 1–19.

Papineau, David (1987). *Reality and Representation*. Blackwell.

Papineau, David (2001). The status of teleosemantics, or how to stop worrying about swampman. *Australasian Journal of Philosophy* 79 (2):279–89.

Perry, John (1980). A problem about continued belief. *Pacific Philosophical Quarterly* 61 (4):317.

Putnam, Hillary (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7:131–93.

Rahimi, Ali (2017). NIPS 2017 Test of Time Award 'Machine learning has become alchemy'. https://www.youtube.com/watch?v=x7psGHgatGM.

Rashid, Tariq (2016). *How To Make Your Own Neural Network*, Createspace Independent Publishing Platform.

Raz, Joseph (1975). *Practical Reason and Norms*. Hutchinson.

Recanati, François (2012). *Mental Files*. Oxford University Press.

Rescorla, Michael (2015/20). The computational theory of mind. *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind>.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–44). ACM.

Russell, Bertrand (1905). On denoting. *Mind* 14 (56):479–93.

Salmon, Nathan U. (1986). *Frege's Puzzle*. Ridgeview.

Soames, Scott (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.

Scanlon, Thomas (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Schubbach, Arno (forthcoming). Judging machines. Philosophical aspects of deep learning. *Synthese* 1–21, available at <https://link.springer.com/article/10.1007/s11229-019-02167-z>

Searle, John R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Searle, John R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (3):417–57.

Shea, Nicholas (2018). *Representation in Cognitive Science*. Oxford University Press.

Soames, Scott (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford University Press.

Soames, Scott (2015). *What Is Meaning*? Princeton University Press.

Soames, Scott (2019). Propositions as cognitive acts. *Synthese* 196 (4):1369–83.

Strawson, P.F. (1959). On referring. *Mind* 59 (235):320–44.

Tam, Kar Yan (1991). Neural network models and the prediction of bank bankruptcy. *Omega* 19 (5):429–45.

Williams, J. Robert G. (2005). *The Inscrutability of Reference,* PhD dissertation, University of St Andrews.

Williams, J. Robert G. (2007). Eligibility and inscrutability. *Philosophical Review* 116 (3):361–99.

Williamson, Timothy (2007). *The Philosophy of Philosophy*. Oxford University Press.

Zou, James and Schiebinger, Londa (2018). AI can be sexist and racist—it's time to make it fair, Nature. Retrieved from: https://www.nature.com/magazine-assets/d41586-018-05707-8/d41586-018-05707-8.pdf.