AI and Human Behavior

Instructor and developer: Prof. Uri Hertz, uhertz@cog.haifa.ac.il , uhertz@sas.upenn.edu

This course was developed for the Masters in Behavior and Decision Making (MBDS) program at University of Pennsylvania, Fall 2024.

It is aimed at a mixed audience of students with no or little technical background, mostly from social sciences bachelor's degrees.

Not all references were required reading, as some were very technical and used just for introduction of main ideas.

The students on this course had to discuss the weekly topic on the course forum, work together and present and analyze an AI product in class, and write a final paper.

This is a work in progress, both since it was given only once and will probably change every time I will give this course, and as new developments, products and societal issues keep emerging.

Course Description

Overview

Artificial intelligence (AI) holds the promise of providing many services originally carried by humans, from generating art, planning vacations, and providing mental help. While developers and enthusiasts emphasize the power and usefulness of these technologies, promoting their integration into multiple fields of human experience, others point out challenges and dangers of these AI agents. These challenges are not just technical limitation, but as these algorithms become bigger and more complex, we lose our ability to understand and interpret how these models operate and make decisions, and how these are aligned with human users' goals and values. The tools used by behavioral researchers and psychologists to understand human behavior and cognition are in a unique position to provide insights into machine behavior.

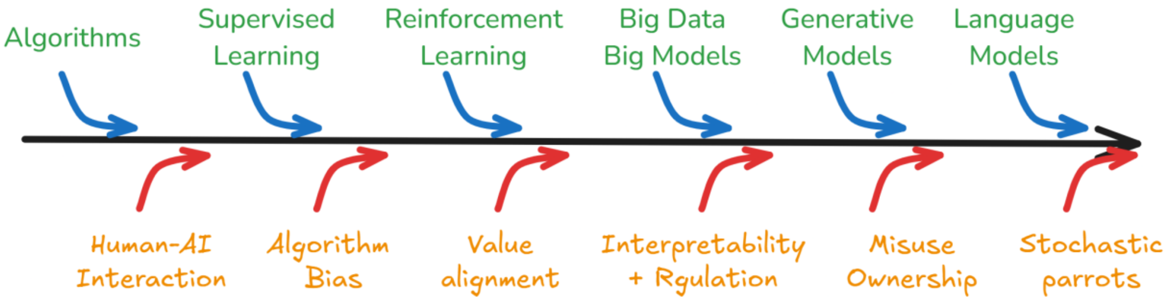
In this course we aim to understand how AI models operate and make decisions, especially when replacing a human decision maker. Throughout the course students will develop an understanding of what are artificial agents, the difference between them and human decision makers, and the challenges and promises they hold.

Goals

- Understand basic definitions and operation of machine learning algorithms.
- Evaluate differences in the way AI and humans perform cognitive tasks.
- Identify factors that shape human and AI interaction.
- Identify social and ethical implications of AI usage and prevalence.

Lecture topics

The course follows two streams, one is technical, where we discuss the architecture and mechanisms of different AI and machine learning models, and the other is about the implications of these models in human society. The streams are interleaved, with different societal issues following a technical discussion.



Below is a detailed description of the lectures, the main points we discussed, and references and sources for each lecture. I don't provide the detailed slides, but feel free to contact me if you want them.

Lecture 0

We discussed how AI and machine learning are general purpose technology (GPT) like the steam engine.

Lecture 1 - Algorithms

What are algorithms? When were they first introduced?

We follow the historic trajectory of algorithms based on Daston, L. (2022). We discuss thin and thick rules, shallow and deep rules, and why algorithms were invented, their relation to automation, context invariance and noise reduction.

We played a drawing game, where one student gives instructions to another student (painter) to draw a picture of a house/face/dog. The painter must try their best not to produce the correct drawing without breaking the rules. This was inspired by Tomer Ullman's loophole works. This illustrated how frustratingly annoying are algorithms, and how different they are from humans.

Daston, L. (2022). Rules: A short history of what we live by. Princeton University Press. https://doi.org/10.1515/9780691239187

Bridgers, S., Schulz, L., & Ullman, T. D. (2021). Loopholes, a Window into Value Alignment and the Learning of Meaning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/content/qt3q48s73j/qt3q48s73j.pdf?t=qwi318

Lecture 2 – Humans and Algorithms

We discussed evidence of algorithmic avoidance, where people prefer not to use algorithms, and algorithmic appreciation.

We talked about different ways to measure these, and how these tendencies may depend on domain, such as financial decisions, medical and moral decisions. We also discussed human-machine interactions in social domains.

We discussed Morewedge's human preference framework.

We had a class discussion about which decisions students feel comfortable with delegating to machines, and what type of decisions they prefer to be made by humans.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59–64.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. Science, 359(6373), 325–329.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 21–34.

Bigman, Y. E., & Gray, K. (2020). Life and death decisions of autonomous vehicles. Nature, 579(7797), E1–E2.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 351–367).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

El Zein, M., Bahrami, B., & Hertwig, R. (2019). Shared responsibility in collective decisions. Nature Human Behaviour, 3(6), 554–559.

Gazit, L., Arazy, O., & Hertz, U. (2023). Choosing between human and algorithmic advisors: The role of responsibility sharing. Computers in Human Behavior: Artificial Humans, 100009,

100009.

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. iScience, 24(6), 102679.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.

Mahmoodi, A., Bahrami, B., & Mehring, C. (2018). Reciprocity of social influence. Nature Communications, 9(1), 2474.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. University of Minnesota Press.

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2022.07.007

Lecture 3 – Machine Learning

We introduced the perceptron, its historical origins and followed a detailed example of training algorithm. We also discussed decision trees, and how to generate them, and the difference between these models.

We played the party game – A person throws a party and comes up with a rule about who is allowed at the party. Each other player tells what gift they brought, and the party owner says if the gift meets the criteria. Everyone needs to guess what the rule was.

Lindsay, G. (2021). *Models of the mind: how physics, engineering and mathematics have shaped our understanding of the brain*.

Lecture 4 – Algorithmic Bias

We talked about different sources of algorithmic bias, and how they relate to the specific nature of machine learning. We discussed the framework proposed by Friedman and Nissenbaum of preexisting bias, technical bias and emergent bias, and its more detailed version that includes labeling, implementation and other problems by van Giffen et al.

Students had to think about specific product, and give example to potential bias, following van Giffen.

Friedman ,B & ,.Nissenbaum, H. (1996). Bias in computer systems *.ACM Transactions on Information Systems*.347–330 ,(3)*14* ,

Guilbeault, D., Delecourt, S., Hull, T., Desikan, B. S., Chu, M., & Nadler, E. (2024). Online images amplify gender bias. Nature, 626(8001), 1049–1055.

van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93–106.

Lecture 5 – Reinforcement Learning

We introduced basic ideas of reinforcement learning, and terms like state-action, value, prediction error, and updating rate. We talked about Skinner, Rescorla -Wagner, and learning in humans, animals and machines.

We used a two-armed bandit example here https://sdl-exp.site/TwoArmsTut/ and discussed how we learned which door is better, what drives this learning and how similar it is to perceptron (rewards instead of labels).

Lindsay, G. (2021). *Models of the mind: how physics, engineering and mathematics have shaped our understanding of the brain*.

Lecture 6- Value-Alignment

We discussed value alignment in humans, e.g., why politicians misrepresent their constituency's values, following Kerr 1975.

We introduced the notion that algorithmic agents may misunderstand and misrepresent human values, as these are introduced as proxies through reward (and cost functions more broadly) and labels.

We discussed algorithmic deceit, reward gaming, and introducing human-in-the-loop training.

We played a game inspired by Ho et al., where the students needed to specify the rewards and punishments of each movement in a 3X3 grid, to make an agent get from one corner to the other without stepping on some squares (a flower patch). We discussed how one should specify a reward policy, and how agents may hack this policy.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2019). People teach with rewards and punishments as communication, not reinforcements. Journal of Experimental Psychology. General, 148(3), 520–549.

Kerr, S. (1975). On the Folly of Rewarding A, While Hoping for B. Academy of Management Journal, 18(4), 769–783.

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O. P., Williams, D., Campbell-Gillingham, L., Thacker, P., Botvinick, M., & Summerfield, C. (2022). Human-centred mechanism design with Democratic AI. Nature Human Behaviour, 6(10), 1398–1407.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). AI Safety Gridworlds. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/1711.09883.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns (New York, N.Y.), 5(5), 100988.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782), 350–354.

Lecture 7 – Big Data and big models

We discussed how big datasets transformed machine learning, the story of ImageNet and AlexNet. We introduced the notion of deep neural networks, and how they improve on perceptrons, and the mechanism of backpropagation. We also talked about overfitting, extrapolation and interpolation and generalization. We compared human and algorithmic face perception and classification.

We played a labeling game – students did a small labeling task online, where they gave labels – one word and a sentence – to multiple images of animals. We discussed the variability of these labels, how people converge, and the features that may shape them. We compared them to labels generated by algorithm.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron, 105(3), 416–434.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science, 332(6025), 60–65.

Huth, A., de Heer, W., Griffiths, T. Theunissen, Gallant . Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458 (2016). https://doi.org/10.1038/nature17637

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? Proceedings. Biological Sciences, 285(1888), 20181319

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for ModelingBiological Vision and Brain Information Processing. Annual Review of Vision Science, 1(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. Trends in Cognitive Sciences, 22(9), 794–809.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by backpropagating errors. Nature, 323(6088), 533–536.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In arXiv [cs.CV]. arXiv.

Young, A. W., & Burton, A. M. (2018). Are we face experts? Trends in Cognitive Sciences, 22(2), 100–110.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3), 107–115.

Lecture 8 - Interpretability

We discussed the black box problem, what are explanations, and the way humans use models (theory of mind) to explain and predict the behavior of other entities (folk psychology). We then discussed how AI models lend themselves to such intuitive explanations, and how these are actually wrong.

We talked about the importance of interpretability for usage, safety and regulation, and the challenges and solutions for interpretability. We mostly stressed how important it is not to assume that our intuitive explanations are correct.

Students had to pick an AI product and provide their intuitive explanation of how this product works. We then highlighted whether these explanations are accurate or not, and reminded ourselves how ML and RL operate.

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. AI and Ethics, 2(3), 377–387.

Cohen, M. K., Kolt, N., Bengio, Y., Hadfield, G. K., & Russell, S. (2024). Regulating advanced artificial agents. Science (New York, N.Y.), 384(6691), 36–38.

Ding, J., Condon, A. & Shah, S.P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun 9, 2002 (2018). https://doi.org/10.1038/s41467-018-04368-5

Fintz, M., Osadchy, M., & Hertz, U. (2022). Using deep learning to predict human decisions and using cognitive models to explain deep learning models. Scientific Reports, 12(1), 4736.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80–89.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. The American Journal of Psychology, 57(2), 243.

Huff DT, Weisman AJ, Jeraj R. Interpretation and visualization techniques for deep learning models in medical imaging. Phys Med Biol. 2021 Feb 2;66(4):04TR01. doi: 10.1088/1361-6560/abcd17. PMID: 33227719; PMCID: PMC8236074.

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. Advances in Neural Information Processing Systems, Nips, 2288–2296.

Lai, X., & Patrick Rau, P.-L. (2021). Has facial recognition technology been misused? A public perception model of facial recognition scenarios. Computers in Human Behavior, 124(106894), 106894.

Lee, T. H., & Boynton, L. A. (2017). Conceptualizing transparency: Propositions for the integration of situational factors and stakeholders′ perspectives. Public Relations Inquiry, 6(3), 233-251. https://doi.org/10.1177/2046147X17694937

Marañes, C., Gutierrez, D. & Serrano, A. Revisiting the Heider and Simmel experiment for social meaning attribution in virtual reality. Sci Rep 14, 17103 (2024). https://doi.org/10.1038/s41598-024-65532-0

Nass, C., Moon, Y., & Carney, P. (1999). Are people polite to computers? Responses to computer-based interviewing systems1. Journal of Applied Social Psychology, 29(5), 1093–1109.

Nussberger, A.-M., Luo, L., Celis, L. E., & Crockett, M. J. (2022). Public attitudes value interpretability but prioritize accuracy in Artificial Intelligence. Nature Communications, 13(1), 5821.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior, 98, 277–284.

Smith, M., & Miller, S. (2022). The ethical application of biometric facial recognition technology. AI & Society, 37(1), 167–175.

Thornton, A., & McAuliffe, K. (2006). Teaching in wild meerkats. Science (New York, N.Y.), 313(5784), 227–229.

Hoppitt, W. J., Brown, G. R., Kendal, R., Rendell, L., Thornton, A., Webster, M. M., & Laland, K. N. (2008). Lessons from animal teaching. Trends in ecology & evolution, 23(9), 486-493.

Lecture 9 – Generative Models

We introduced the notion of algorithms that generate data, and how these generated stimuli are different from real world data, revisiting the 'fit to nature' paper. We talked about encoder-decoder architecture and image processing (inpainting, pattern removal), generative adversarial networks (GANs), and diffusion processes. We highlighted the new modular nature of these models, and how they combine different components to expand their capacity, especially language and image creation.

Students also tried different generation models in class (images and text). They asked models to reproduce lyrics for songs based on their title, mostly non-USA artists, recipes for dishes, and images of places and events from their own culture. We discussed the difference between real events and images, and generated ones. We especially discussed how generative models create *plausible* stimuli, but do not recover real data.

Cai, N., Su, Z., Lin, Z. et al. Blind inpainting using the fully convolutional neural network. Vis Comput 33, 249–261 (2017). https://doi.org/10.1007/s00371-015-1190-z

Dotsch, R., & Todorov, A. (2012). Reverse Correlating Social Face Perception. Social Psychological and Personality Science, 3(5), 562–571.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1406.2661

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron, 105(3), 416–434.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In arXiv[cs.LG]. arXiv. http://arxiv.org/abs/2006.11239

Jiao, W., Atwal, G., Polak, P. et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun 11, 728 (2020). https://doi.org/10.1038/s41467-019-13825-8

Zhang, X., Zhai, D., Li, T., Zhou, Y., & Lin, Y. (2023). Image inpainting based on deep learning: A review. An International Journal on Information Fusion, 90, 74–94.

Lecture 10 – Productivity, Creativity, and Reality

We discussed three ways in which generative models affect human society. We started by discussing productivity, and how these tools can automate many procedures in flexible ways. We then discussed creativity, and whether models make us more creative or homogenize human outcomes. Finally, we discussed the way these models can mimic and generate reality, and the implications on epistemic trust. We compared these models to the introductions of journalism, the camera, videos and other means that could bring far away events, beyond our immediate environment and sense, to millions of people.

We played with creative models, trying to evaluate how original or satisfying their output was, and whether these can serve as a final product.

Almeida, D., Shmarko, K., & Lomas, E. (2022). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. AI and Ethics, 2(3), 377–387.

Anderson, B. R., Shah, J. H., & Kreminski, M. (2024, June 23). Homogenization effects of large language models on human creative ideation. Creativity and Cognition. C&C ′ 24: Creativity and Cognition, Chicago IL USA. https://doi.org/10.1145/3635636.3656204

Anil R. Doshi, Oliver P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci. Adv.10,eadn5290(2024).DOI:10.1126/sciadv.adn5290

Anstine, D. M., & Isayev, O. (2023). Generative models as an emerging paradigm in the chemical sciences. Journal of the American Chemical Society, 145(16), 8736–8750.

Arias-Sarah, P., Bedoya, D., Daube, C., Aucouturier, J.-J., Hall, L., & Johansson, P. (2024). Aligning the smiles of dating dyads causally increases attraction. Proceedings of the National Academy of Sciences of the United States of America, 121(45), e2400369121.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? Trends in Cognitive Sciences, 27(7), 597–600.

Kauppinen, A. (2018). Epistemic Norms and Epistemic Accountability. Philosopher’s Imprint, 18(8). https://quod.lib.umich.edu/p/phimp/3521354.0018.008/1

Koh, E., & Doroudi, S. (2023). Learning, teaching, and assessment with generative artificial intelligence: towards a plateau of productivity. Learning: Research and Practice, 9(2), 109–116. https://doi.org/10.1080/23735082.2023.2264086

Luccioni, S., Jernite, Y., & Strubell, E. (2024, June 3). Power hungry processing: Watts driving the cost of AI deployment? The 2024 ACM Conference on Fairness, Accountability, and Transparency.

Mercier, H. (2020). Not Born Yesterday. Princeton University Press.

Messeri, L., Crockett, M.J. Artificial intelligence and illusions of understanding in scientific research. Nature 627, 49–58 (2024). https://doi.org/10.1038/s41586-024-07146-0

Regenwetter, L., Nobari, A. H., & Ahmed, F. (2022). Deep generative models in engineering design: A review. Journal of Mechanical Design (New York, N.Y.: 1990), 144(7), 071704.

Shakked Noy, Whitney Zhang ,Experimental evidence on the productivity effects of generative artificial intelligence. Science 381,187-192(2023).DOI:10.1126/science.adh2586.

Sperber, D. A. N., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. Mind & Language, 25(4), 359–393.

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief. Trends in Cognitive Sciences, 22(3), 213–224.

Wang, H., Fu, T., Du, Y. et al. Scientific discovery in the age of artificial intelligence. Nature 620, 47–60 (2023). https://doi.org/10.1038/s41586-023-06221-2

Wornow, M., Xu, Y., Thapa, R. et al. The shaky foundations of large language models and foundation models for electronic health records. npj Digit. Med. 6, 135 (2023). https://doi.org/10.1038/s41746-023-00879-8

Lecture 11 – Large Language Models

We discussed script based and symbolic based NLPs (and Eliza), as well as statistical approaches like bag-of-words. We discussed word2vec, and the concept of attention and transformers. We discussed GPT and BERT as flexible modular models.

We emphasized the notion that LLMs are at heart the same as AlexNet for object detection, but the statistical landscape they capture is language, whose statistical regularities follow dimensions like semantics, reasoning, and culture. This makes them feel like humans, because humans are the only creatures we know that use language in this way.

In class we played with translation of single words and sentences.

Bhatia, S. (2017). Associative judgment and vector space semantics. Psychological Review, 124(1), 1–20.

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Havron, N., de Carvalho, A., Fiévet, A.-C. and Christophe, A. (2019), Three- to Four-Year-Old Children Rapidly Adapt Their Predictions and Use Them to Learn Novel Word Meanings. Child Dev, 90: 82-90. https://doi.org/10.1111/cdev.13113

Lior, G., Shalev, Y., Stanovsky, G., & Goldstein, A. (2024). Computation or Weight Adaptation? Rethinking the Role of Plasticity in Learning. bioRxiv, 2024-03.

Malmaud, J., Levy, R., & Berzak, Y. (2020). Bridging information-seeking human gaze and machine reading comprehension. arXiv preprint arXiv:2009.14780.

Mercier, H., & Sperber, D. (2017). The Enigma of Reason (p. 396). Harvard University Press.

Radford, A. (2018). Improving language understanding by generative pre-training.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1706.03762

Lecture 12 – Intelligence, Benchmarks and AGI

We discussed the notion of intelligence, the Turing test, and benchmark approach to testing intelligence (ARC). We went over a number of fallacies of LLMs, and played a bit with LLMs, examining how these were solved using patches and methods like train-of-thoughts and reasoning. We also noted how the current efforts towards AGI are less ML techniques, and more the use of heuristics and intuitions from cognitive psychology. We read Melanie Mitchel's blog posts on AI.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March 3). On the dangers of stochastic parrots: Can language models be too big? . Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2023). The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A." International Conference on Learning Representations, abs/2309.12288. https://doi.org/10.48550/arXiv.2309.12288

Bhatia, S. (2024). Exploring variability in risk taking with large language models. Journal of Experimental Psychology. General, 153(7), 1838–1860.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences of the United States of America, 120(6), e2218523120.

Chollet, F. (2019). On the measure of intelligence. In arXiv [cs.AI]. arXiv. http://arxiv.org/abs/1911.01547

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. In arXiv [cs.AI]. arXiv. http://arxiv.org/abs/1803.05457

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625–630.

Gray, K., Yam, K. C., Zhen'An, A. E., Wilbanks, D., & Waytz, A. (2023). The psychology of robots and artificial intelligence. The Handbook of Social Psychology. https://www.deepestbeliefslab.com/s/psychology-of-robots-and-ai.pdf

Koriat, A. (2012). When are two heads better than one and why? Science , 336(6079), 360–362.

Levinstein, B. A., & Herrmann, D. A. (2024). Still no lie detector for language models: probing empirical and conceptual roadblocks. Philosophical Studies, 1–27.

Mitchell, M. (2024). The metaphors of artificial intelligence. Science (New York, N.Y.), 386(6723), eadt6140.https://www.science.org/doi/full/10.1126/science.adt6140?af=R

Nezhurina, M., Cipolina-Kun, L., Cherti, M., & Jitsev, J. (2024). Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. arXiv preprint arXiv:2406.02061.

Simchon, A., Edwards, M., & Lewandowsky, S. (n.d.). The persuasive effects of political microtargeting in the age of generative AI. PNAS Nexus. https://doi.org/10.1093/pnasnexus/pgae035

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing theory of mind in large language models and humans. Nature Human Behaviour, 1–11.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv preprint arXiv:2309.13638.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

Yax, N., Anlló, H., & Palminteri, S. (2024). Studying and improving reasoning in humans and machines. Communications Psychology, 2(1), 1–16.

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission Versus Truth, Imitation Versus Innovation: What Children Can Do That Large Language and Language-and-Vision Models Cannot (Yet). Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 17456916231201401.