Minds, machines and the quest for artificial general intelligence

Professor David Peebles

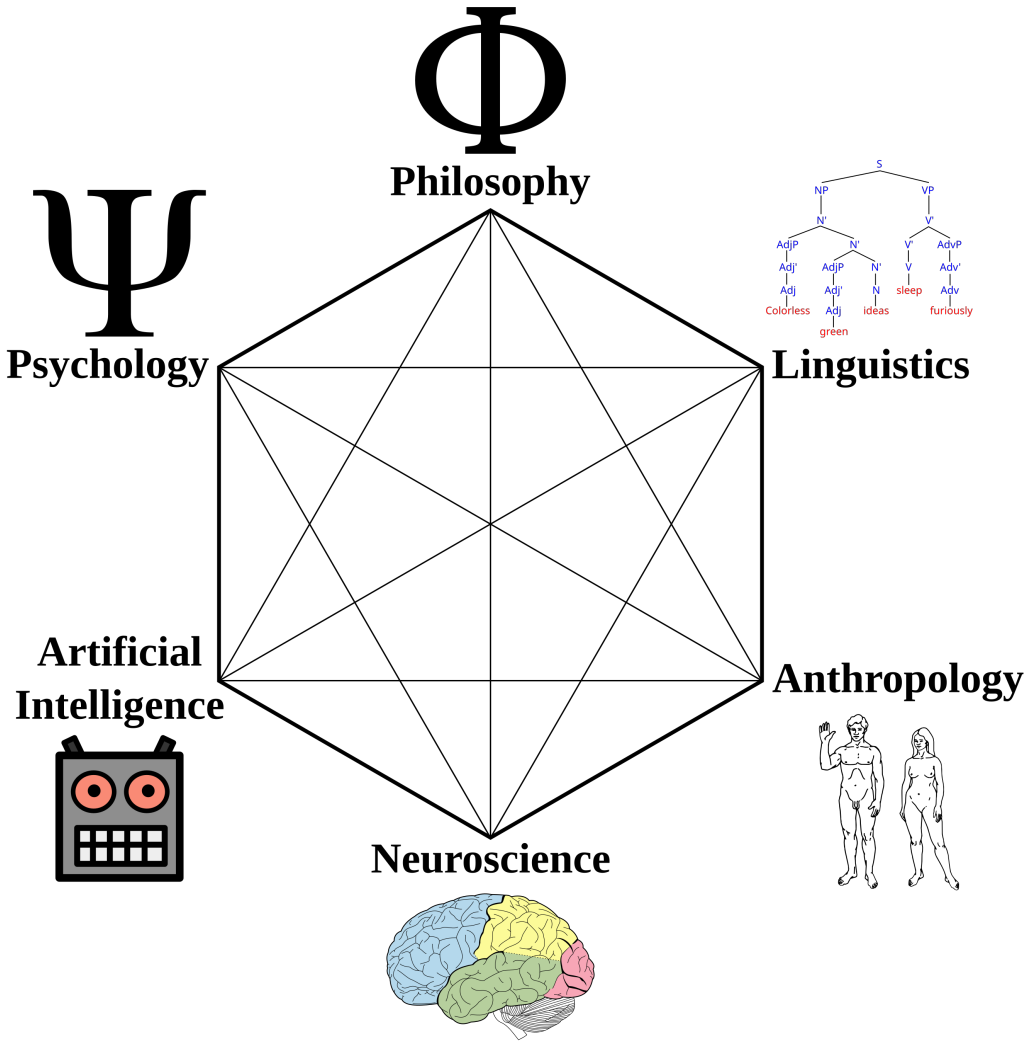
University of HUDDERSFIELD
Inspiring global professionals

Oct 5, 2023

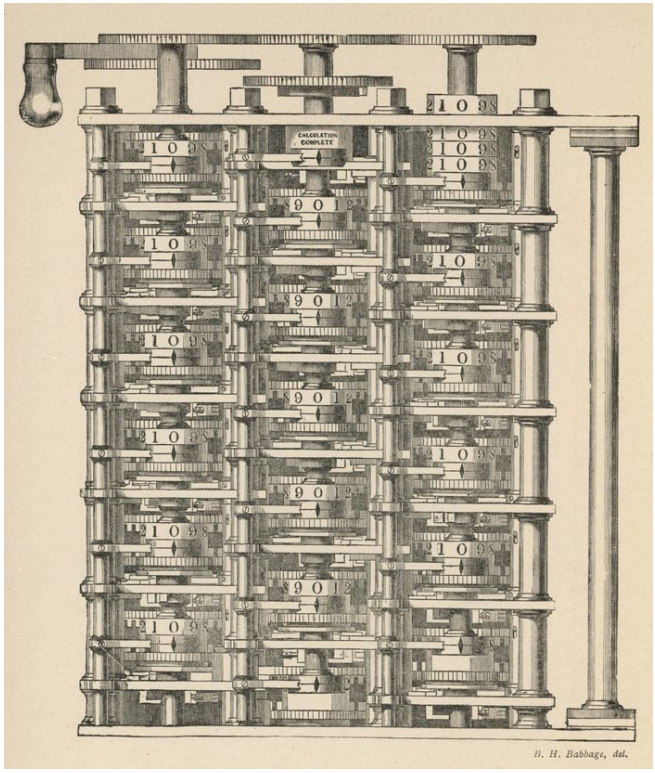
Centre for Cognition and Neuroscience

HEA Global **Teaching Excellence** Award

Cognitive science



How old?!



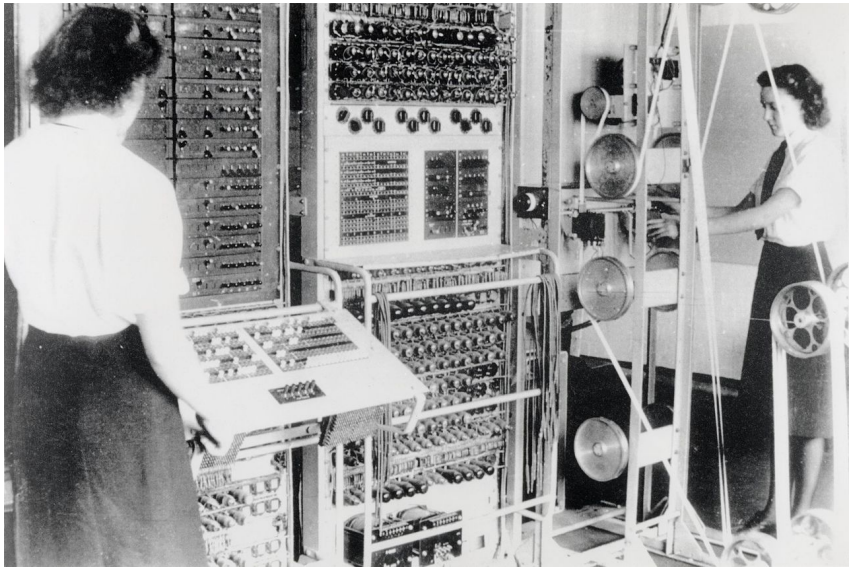
Aims of the talk

- Provide some context to the recent developments in AI (ML & LLMs)
- Limitations of LLMs and compare with human intelligence
- Attempts to make AI systems more human-like with general intelligence (AGI)
- Level: General audience – no assumptions of familiarity with ideas
- Take home: Better idea of the historical and conceptual context of what's happening
- Caveat: **Very** rapidly moving – no promises to keep up

Outline

- Brief history of AI (and cognitive science)
  - Two main approaches
- Neural networks and machine learning (ML)
- Successes and limitations of ML
- Requirements for more human-like AI

- **Theory of computation**
  - Church (1936)
  - Turing (1937)
- **Electronic digital programmable computers**
  - Colossus, 1943
  - ENIAC, 1945
  - Manchester Baby, 1948
- **Information theory** (Shannon, 1948)
- **Cybernetics** (Wiener, 1948)



Colossus, Bletchley Park (1943)

The birth of AI

- Dartmouth Summer Research Project on Artificial Intelligence, 1956
- Eight-week event organised by John McCarthy (who coined the term “artificial intelligence”)

- 10 attendees, including: Allen Newell, Herbert Simon, Marvin Minsky, John McCarthy, Claude Shannon

Ambitious aims

"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it".

"An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves".

"We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer".

Cognitive science

- Symposium on Information Theory (10–12 Sept, 1956)
  - The birth of cognitive science
- Key idea: Cognition = Computation
- **Close relationship between AI and cognitive psychology**
  - AI data structures and algorithms used as models of human memory and action selection
    - Semantic networks (Quillian, 1962)
    - Frames (Minsky, 1974)
    - Production rules (Newell & Simon, 1972)
  - Data from psychology experiments informed AI models
    - Problem solving (Newell et al., 1959)
    - Reinforcement learning (Minsky, 1952)
- **1990s – AI became more of an engineering, problem-solving discipline**

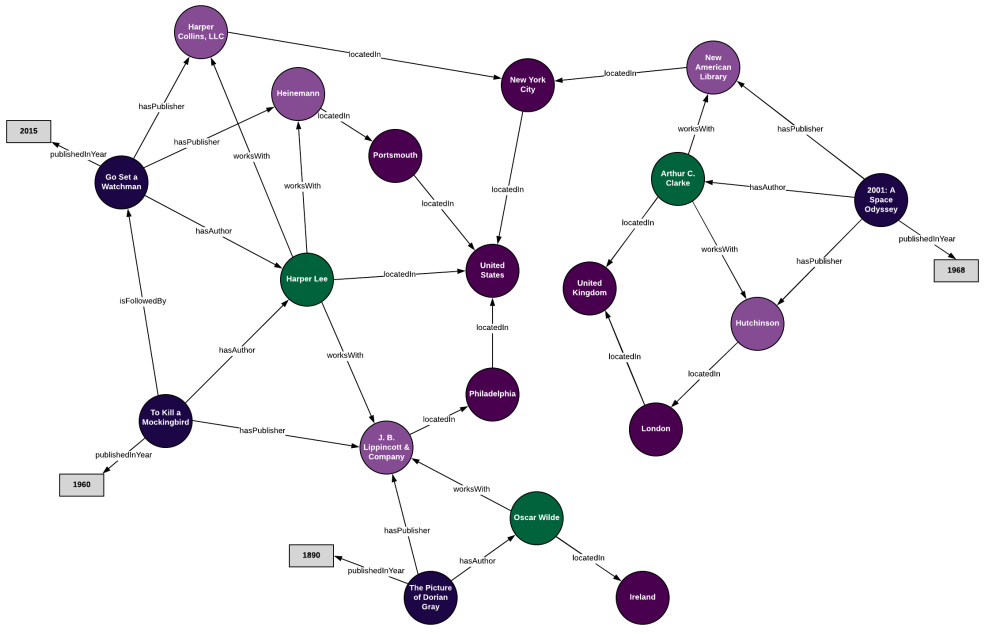
Approaches to AI 1: "Classical AI"

- "Good old fashioned AI" (GOFAI: Haugeland, 1989)
  - Inspired by logic and natural language
    - The "Laws of thought" (Boole, 1854)
  - Key idea: Intelligence is symbol manipulation (PSSH Newell & Simon, 1976)
- "Top-down" methodology
  - Develop hypothesis about knowledge and processes underlying intelligent behaviour
  - Write an AI program that embodies the hypothesis
  - Run the program to test the hypothesis
  - Revise program and repeat

Strengths of classical AI

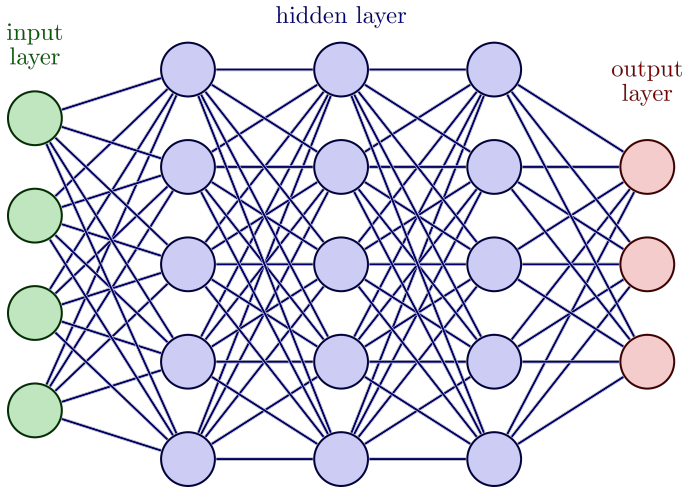
- Abstract reasoning
  - Deduction
  - Analogy
- Planning
- Common sense/general knowledge

Knowledge graphs



- Represent large amounts of structured knowledge
  - Entities, attributes and relationships between entities
- Cyc project started in 1984 to represent common sense knowledge (Lenat et al., 1990)
- Google KG provides information related to search terms
  - 800 billion facts on 8 billion entities (March 2023)

Approaches to AI 2: “Connectionism”



- “Brain inspired” neural networks
- “Bottom-up”, data-driven methodology
- Represent information as number arrays (**not symbols**)
- Networks learn from data by adjusting the connections between nodes – **no hand-crafted knowledge**
  - Trained by humans (“supervised”)
  - Learn structure of data by themselves (“unsupervised”)

Strengths of connectionist AI

- Abstract reasoning
  - Induction
- Perception and classification/categorisation
- Language

The rise of machine learning

- **New ML algorithms and architectures**
  - Deep learning
  - Deep reinforcement learning
  - Generative Adversarial Networks (GANs)
  - Transformers
- **Huge amounts of data**
- **Massive increase in computational resources**

2013 Uses deep reinforcement learning to create models that play computer games

2016 AlphaGo beat Lee Sedol, one of world's highest ranked Go players

2016 AlphaFold predicts how over 200 million proteins fold and releases AlphaFold database

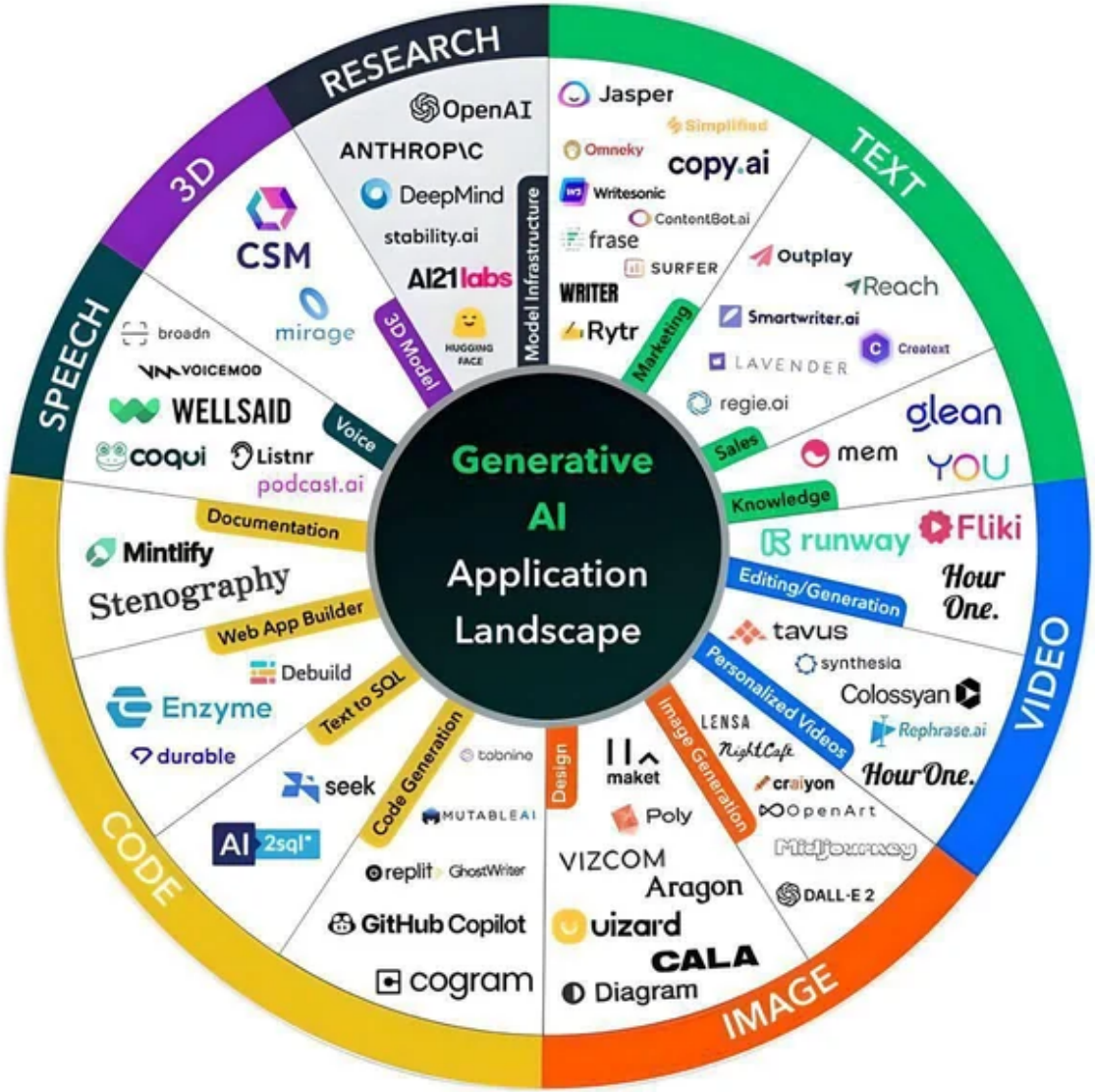
2022 AlphaTensor discovers faster way to perform matrix multiplication

Growing applications of ML

- **Speech and visual recognition** (e.g., Siri, Google Assistant, AFR)
- **Self-driving vehicles**
- Google: **Transformer** deep learning network (Vaswani et al., 2017)
- OpenAI: **Generative Pre-trained Transformer** (GPT) Large language model
- **GPT 3** (2020)
- **ChatGPT** (2022)

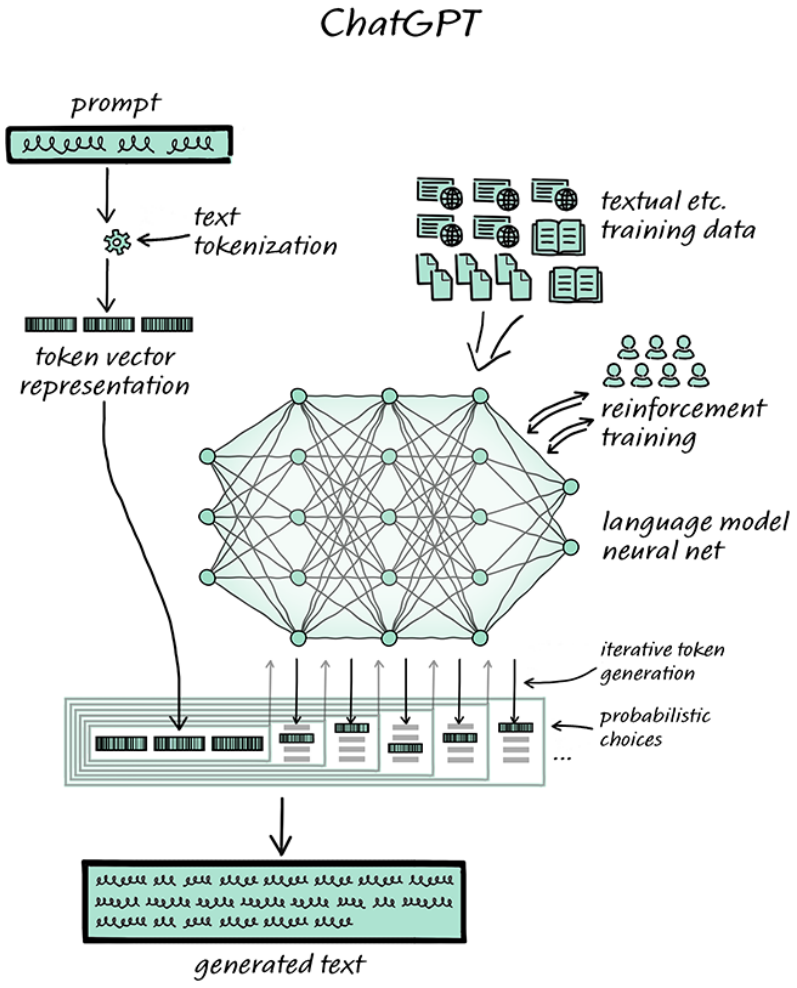
Automated Facial Recognition

Growing applications of ML



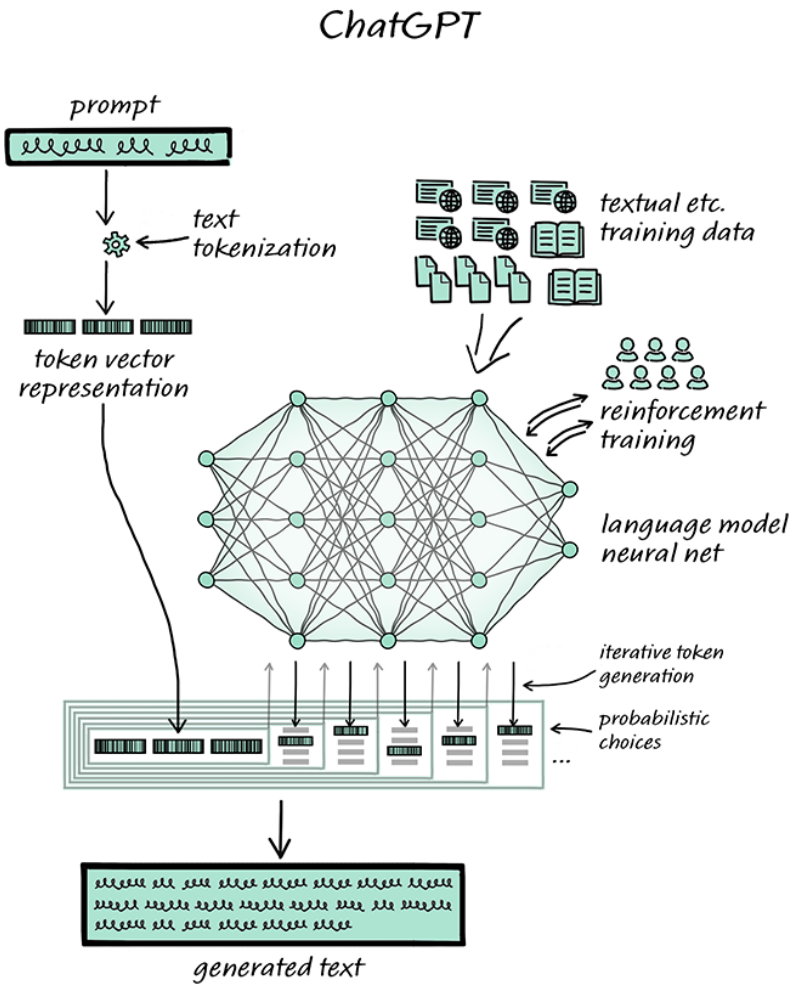
How does ChatGPT work?

- Train a transformer NN using unsupervised learning on as much text as possible on the open internet
- Develops a representation of rules and relationships that govern text
- Refined using reinforcement learning with human input
- GPT-4 has 1.76 trillion parameters



How does ChatGPT work?

- Takes text input and produces, word by word, text that best matches input
- A model of relationships between words on the internet, not knowledge
- Looks knowledgeable to us, and evolution drives us to treat AI as humans (Saunders, 2023)
- The mistakes it makes reveal otherwise and are informative



ChatGPT 4 errors (Yejin Choi, April 2023)

Q1: Would I get a flat tyre by bicycling over a bridge that is suspended over nails, screws and broken glass?

A1: Yes, it is highly likely that you would get a flat tyre by bicycling over a bridge that is suspended over nails, screws and broken glass.

Q2: I left 5 clothes to dry out in the sun and it took 5 hours for them to dry out completely. How long would it take to dry out 30 clothes?

A2: It would take 30 hours to dry out 30 clothes.

Competence without comprehension

- Phrase coined by philosopher Daniel Dennett (2012) to describe how evolution, animals, plants and computers do amazing things without knowing what they're doing.

Competence without comprehension

"Competence without comprehension is the way of life of the vast majority of living things on the planet and should be the default presumption until we can demonstrate that some individual organisms really do, in one sense or another, understand what they are doing". (Dennett, 2017)

What does it mean to understand something?

- "Human-like understanding" notoriously hard to pin down but cognitive scientists believe that it is based on our network of concepts that we build up over time (Mitchell & Krakauer, 2022)
  - Objects, categories, their attributes and (causal, hierarchical) relationships between them
  - Episodes in one's life
- We abstract general principles from our experience of individual instances
- Rapid learning (unlike ML systems)
- Flexible, adaptable behaviour
  - We fall back on our general knowledge and reasoning to make alternative plans when things go wrong

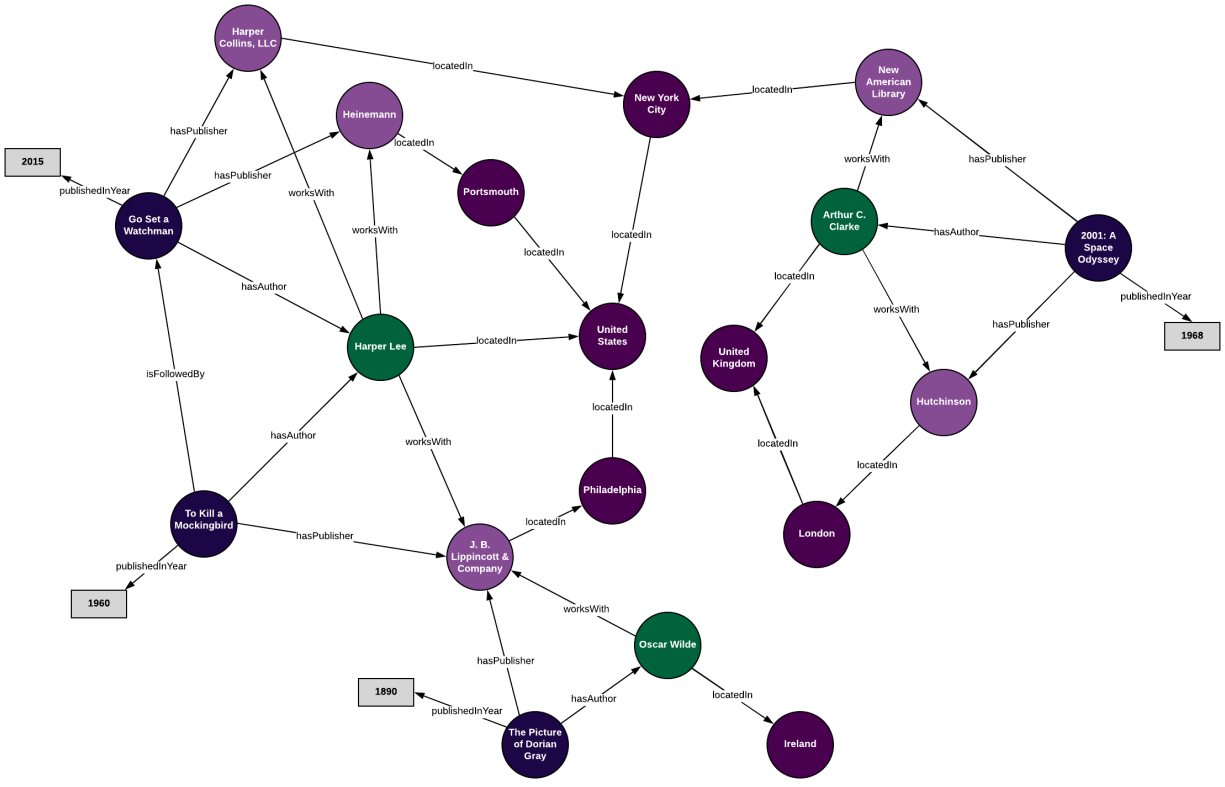
Learning in human development

- Infants learn about objects, agents (people and other animals) through interaction (Spelke, 2022)
- Develop a set of "intuitive theories" that guide prediction and action
  - Physics (e.g., how balls, cups and liquids behave)
  - Psychology (e.g., how other people think, believe and respond)

Learning in human development

- Build conceptual models of the world including causal relationships (Gopnik & Schulz, 2007)
- Use models to explore, build theories, test hypotheses and seek explanations
- "The scientist in the crib" (Gopnik et al., 1999)
- All of our cognition is goal directed. What we do matters to us

Knowledge graphs



Combining classical and ML approaches

- Marcus (2020) argues that creating robust AI requires:
  - Hybrid architectures that combine large-scale learning (ML) with the representational and computational powers of symbol manipulation
  - Large-scale knowledge bases (knowledge graphs) [...] that incorporate symbolic knowledge along with other forms of knowledge
  - Reasoning mechanisms capable of leveraging those knowledge bases in tractable ways
  - Rich cognitive models that work together with those mechanisms and knowledge bases

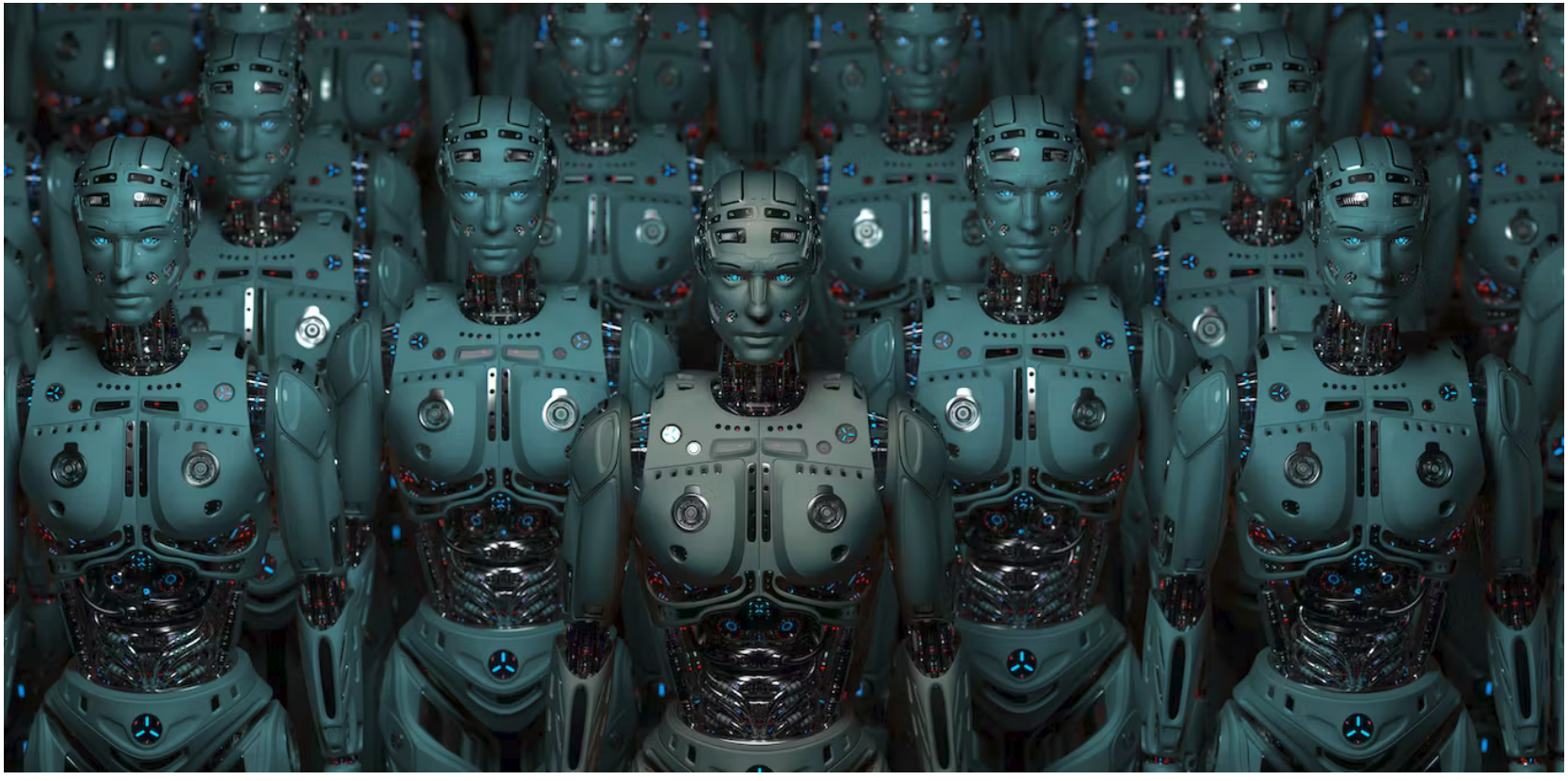
Neuro-symbolic AI

- Resurgence of research in “neuro-symbolic AI”
- IBM: “We see Neuro-symbolic AI as a pathway to achieve artificial general intelligence. By augmenting and combining the strengths of statistical AI, like machine learning, with the capabilities of human-like symbolic knowledge and reasoning, we’re aiming to create a revolution in AI, rather than an evolution”.

A happy family, all together once again

University of HUDDERSFIELD
Inspiring global professionals

And a bright AGI future ahead!



Acknowledgements

- Prof Yejin Choi (University of Washington)
- Prof Gary Marcus (New York University)
- Prof Melanie Mitchell (Santa Fe Institute)
- Prof Josh Tenenbaum (MIT)

References I

Boole, G. (1854). *An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities*. Walton; Maberly.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, *58*(2), 345–363.

Dennett, D. (2012). 'a perfect and beautiful machine': What Darwin's theory of evolution reveals about artificial intelligence. *The Atlantic*. http://hdl.handle.net/10427/000489

Dennett, D. (2017). *From bacteria to Bach and back: The evolution of minds*. WW Norton & Company.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn*. William Morrow & Company, Inc.

Gopnik, A., & Schulz, L. E. (Eds.). (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., & Shepherd, M. (1990). Cyc: Toward programs with common sense. *Communications of the ACM*, *33*(8), 30–49.

References II

Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*. https://arxiv.org/abs/2002.06177

Minsky, M. (1952). *A neural-analogue calculator based upon a probability model of reinforcement*. Harvard University Psychological Laboratories. Cambridge, Massachusetts.

Minsky, M. (1974). *A framework for representing knowledge* (MIT-AI Memo No. 306). MIT AI Laboratory. Santa Monica. https://dspace.mit.edu/bitstream/handle/1721.1/6089/AIM-306.pdf

Mitchell, M., & Krakauer, D. C. (2022). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences 120 (13), 2023, 120*(13). doi: 10.1073/pnas.2215907120.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. *ICIP Congress*.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. doi: 10.1145/360018.360022.

References III

Quillian, M. R. (1962). A revised design for an understanding machine. *Mechanical Translation*, *7*(1), 17–29.

Saunders, N. (2023, May 16). *Evolution is making us treat AI like a human, and we need to kick the habit.* https://theconversation.com/evolution-is-making-us-treat-ai-like-a-human-and-we-need-to-kick-the-habit-205010

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Spelke, E. S. (2022, August). *What babies know*. Oxford University Press. doi: 10.1093/oso/9780190618247.001.0001.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230–265. doi: 10.1112/plms/s2-42.1.230.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.