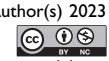
Original Research Article

SCE

Studies in Christian Ethics
2023, Vol. 36(3) 604–619



DOI: 10.1177/09539468231169462
journals.sagepub.com/home/sce


Artificial Intelligence (and Christianity): Who? What? Where? When? Why? and How?

George M. Coghill
Edinburgh Theological Seminary and University of Aberdeen, UK

Abstract

Artificial Intelligence (AI) is a high-profile subject these days. In its brief history it has undergone several highs and lows and suffered from significant degrees of hype as well as antagonism and fear. One thing is clear: we are no closer to the goal of producing a truly sentient being than when it started. Nonetheless, the tools developed by AI researchers are here to stay and as with all technological advances it has its good and bad aspects. In this article I will present a brief overview of the field of AI looking at what it is, how it developed, what are its specialism as well as some of its well-publicised successes, and failures, as well as pointing out some key Christian participants in the story.

Keywords

Artificial Intelligence, Christianity and AI, religion and science, history of AI, history of computing

What is Artificial Intelligence?

When one reads articles, particularly in the Humanities, about Artificial Intelligence (AI) and its impact, one is often left with the impression that the writer considers AI to be a single, coherent, homogeneous body of knowledge/domain of study. In a sense this is inevitable since greater detail and nuance does not serve the writer's purpose. But when that 'homogeneous body' is reduced to Machine Learning (or worse, Deep

Corresponding author:
George M. Coghill, Edinburgh Theological Seminary and University of Aberdeen, UK.
Emails: gcoghill@ets.ac.uk; g.coghill@abdn.ac.uk

Learning) something has gone awry![1] So let me begin by stating that the field of AI is far from homogeneous.

It should be no surprise that there is a multiplicity of approaches to the study of AI given the foundational subjects contributing to and forming the basis of the field. At the very least the following disciplines are relevant: Neuroscience (identifying how the brain processes information); Psychology (explaining how humans think and act); Linguistics (assessing how language relates to thought); Control Theory (analysing how artefacts can operate autonomously); and Computer Science (designing and implementing effective computational reasoning systems).

Given these perspectives, there are several different viewpoints on what constitutes AI and how it should be undertaken. These may be grouped according to whether the primary focus is on humans or rationality, and on whether it is thought or behaviour that is deemed more important.[2]

- *Thinking humanly*: where the aim is to get a machine to think like a human.
- *Thinking rationally*: where the aim is to construct a machine that reasons in the best possible way.
- *Acting humanly*: where the aim is to build a machine that performs in the same manner as humans.
- *Acting rationally*: where the aim is to get a machine to act in the best possible way.

So there is a spectrum of approaches, from Cognitive Science at one end (driven by a desire to understand and model how humans operate) and hard-core engineering at the other (driven by the desire to produce a system that will not be affected by the frailties of human behaviour); because, let's face it, the noetic effects of sin ensure that humans do not always think or behave in a rational manner.

As well as divergence in considering what AI is, there is also a wide variance in how different schools of thought approach the development of AI systems. At one end there are the *formalists*, who insist that in order to be acceptable the system has to be provably correct.[3] At the other are those who recognise that there is much that can be achieved by designing, building and testing systems experimentally, in a more 'messy' way, especially when the issues involved may be too complex to prove (at least in any straightforward manner).[4] All these are represented in the AI community.

Regardless of how one views the whys and wherefores of AI, in order to progress building an AI system one has to have some idea of what needs to go into the system (or agent). This may be as simple as interfacing to instrumentation for data collection, or it may require computer vision

1. Randall Reed, 'AI in Religion, AI for Religion, AI and Religion: Towards a Theory of Religious Studies and Artificial Intelligence', *Religions* 12.6 (2021), p. 401.
2. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd edn (London: Prentice-Hall, 2010).
3. Edsger Dijkstra, 'On the Cruelty of Really Teaching Computing Science' (EWD-1036, 1988), https://www.cs.utexas.edu/~EWD/transcriptions/EWD10xx/EWD1036.html.
4. Guus Schreiber et al., *Knowledge Engineering and Management: The CommonKADS Methodology* (Cambridge, MA: MIT Press, 2000).

and natural language (processing and generating) capabilities. Inside the system what is implemented will depend on what task the agent is required to undertake. In the simplest case it may be just a means of interpreting the input, applying some rules, and performing some action on that basis in the environment. At the other end of the spectrum, an ability to learn in a changing environment is critical for judging the requirements and possibilities relevant to the task at hand (and to change the details of the task if need be). Ideally the system should be able to explain and justify any decisions made or actions taken (though this is not always possible). Each of these elements is a domain of active research, with researchers, in some areas at least, being very focused on their particular problem, rather than AI in general. This focus is simply an indication of the complexity of, even apparently simple, real-world problems.[5]

An Overview of the History of AI

AI as a specialist area of research came to the fore with the invention of the (electronic) computer (analog and digital) in the middle of the twentieth century. However, prior to that there was a history of related efforts that led to that point, and Christians played their part in these developments.

Pre-History

There has always been a strand within humanity that has hankered after the ability to create a being in our own image, with or without *divine* assistance. Literature throughout history is replete with examples. There is the ancient Greek story of Pygmalion who falls in love with a statue he has sculpted and which Aphrodite brings to life for him. Then there is the story of the Golem in the sixteenth century. The Golem is a Jewish myth regarding the animation of, usually, clay.[6] The most famous Golem is the folk tale about Rabbi Judah ben Bezalel of Prague, who made a golem from clay and brought it to life by means of Hebrew rites to protect the Prague Jews from pogroms (a story which finds a place in recent times when Norbert Wiener gives the title *God and Golem, Inc.*,[7] to the popularisation of his technical book *Cybernetics*[8]). But the oldest reference is the Talmudic reference to Adam as starting out as a golem.[9]

A contrasting vision is depicted by Mary Shelley in *Frankenstein*[10] which captures the fearful reaction to the news of Abbé Nolet's experiments stimulating the amputated leg of a frog by electrical means.[11] The subtitle *The Modern Prometheus* more than hints at a

5. Russell and Norvig, *Artificial Intelligence*.
6. Golem is an unformed mass, mentioned in Ps. 139:16 (the only place in the Bible where the word appears).
7. Norbert Wiener, *God and Golem, Inc.* (London: Chapman & Hall, 1964).
8. Norbert Wiener, *Cybernetics: Or the Control and Communication in the Animal and the Machine*, 2nd edn (Cambridge, MA: MIT Press, 1965).
9. Babylonian Talmud, Tractate Sanhedrin 38b. There is also reference in general in Sanhedrin 38a to the unshaped form in Ps. 139:16.
10. Mary Shelley, *Frankenstein* (London: Frontpage Publishing, 2019).
11. John L. Heilbron, 'Nolet, Jean Antoinne', in C.C. Gillespie (ed.), *Dictionary of Scientific Biography* (New York: Charles Scribner & Sons, 1974).

predicted negative outcome for such insolent efforts to 'create' life, that is the sole prerogative of God.[12]

At the beginning of the twentieth century there was Karel Capek's *RUR*[13] (Rosum's Universal Robots), which introduced a new term into the English language that is still relevant to AI (though not all roboticists are interested in AI).[14] Of course, since the inception of AI there have been any number of books covering a variety of issues, starting with Isaac Asimov and his 'Three Laws of Robotics'.[15]

In the world of practical applications, in what Pamela McCorduck called 'The Mechanisation of Thinking'[16] there were also attempts to automate and further formalise the reasoning process. And there was a significant Christian presence in this development. The first of these was the *Ars Magna* of Ramon Llull.[17]

Ramon Llull (c.1232–c.1315) was a medieval mystical Catalan missionary. His early life was secular, but after several mystical experiences, which occurred while he was trying to compose a love poem for a lady other than his wife, he converted to Christianity and spent a significant part of his life as a missionary around the Mediterranean. It was through his interaction with the Islamic world that he came across the *zaijra,* a device used by Arabic astrologers to construct new predictions. From this, Llull constructed a novel approach to logic and reasoning (the *Ars Magna*) that went beyond the syllogistic reasoning of the scholastics, and could propose new findings. He envisaged this as a tool to aid mission work amongst Jews, Muslim and schismatic Christian groups (e.g., Nestorians).[18] Llull had noted that the main difficulty Jews and Muslims had with Christianity was the doctrine of the Trinity. The *Ars Magna*, then, started with principles common to all three religions of the Book, and by means of the then accepted process of the 'ladder of being' making mechanical steps up the ladder ultimately arriving at establishing the Trinity. As such it was a tool used to facilitate discussions between theologians.[19]

Llull's *Ars Magna* has influenced a number of thinkers in the modern era, in particular Gottfried Leibniz.[20] Leibniz (1646–1716) was a Protestant who yearned for the re-unification of the Western church. One major hope was for a time when one would

12. Ironically, the end point of Nolet's experiments is the development of things such as Functional Electrical Stimulation for patients suffering from spinal and other disorders of the nervous system.
13. Karel Capek, *R.U.R.,* trans. Paul Selver and Nigel Playfair (Mineola, NY: Dover Publications, 2001).
14. And the diminutive 'bot' has become a term for any active artificial entity.
15. Isaac Asimov, *I, Robot* (New York: Gnome Press, 1950).
16. Pamela McCorduck, *Machines Who Think*, 2nd edn (London: Routledge, 2004).
17. Thessa Jensen, 'Ramon Llull's *Ars Magna*', in Roberto Moreno-Diaz et al. (eds.), *EUROCAST 2017, Part 1*, vol. 10671 (Cham: Springer International Publishing, 2018).
18. Jensen, 'Ramon Llull's *Ars Magna*'; Ernesto Priani, 'Ramon LLull', in *Stanford Encyclopedia of Philosophy* (Stanford, CA: Stanford University Press, 2021).
19. Making this an early precursor of the interaction between AI and Theology.
20. Gordon H. Clark, *Thales to Dewey: A History of Philosophy*, 4th edn (Unicoi, TN: The Trinity Foundation, 2000).

be able to settle disputes by means of a formal reasoning process that put conclusions beyond doubt.[21]

A more recent contribution from within the Christian traditions relevant to the history of AI is that of Charles Babbage (1791–1871). Babbage is best remembered for his contributions to Computer Science. A 'computer' in the Victorian age was a human employed to calculate navigational tables, a very error-prone process. Babbage's design of the Difference Engine was meant to overcome these deficiencies by automating the process. While his Difference Engine was implemented, the more complex Analytical Engine never saw completion. Babbage contributed to the series of Bridgewater Treatises a work on Natural Theology. This work was cited by Robert Chambers in the famous 'Vestiges'. He considered that Babbage's computational outlook as making it plausible that the transmutation of species could have been pre-programmed.[22]

Babbage did not envisage that mechanical computational devices would be used for anything other than routine calculations. His colleague Ada Lovelace (1815–1852), the world's first systems analyst,[23] had a different view. While she did not envisage anything like General AI, she did foresee a time when machines could be used to perform other, more 'intelligent' tasks along the lines of what we would term narrow AI.[24]

Its History 'Proper'

With the creation of electrical computation machines, and from the 1950s, electronic devices, things started to move quite quickly. In 1943, a collaboration between a physiologist (Warren McCulloch: 1898–1969) and a logician (Walter Pitts: 1923–1969) resulted in a Boolean model of the brain.[25]

As things were starting to develop, approaches to AI formed two strands depending on one's background and computational predilections: analog or digital. The former rose out of control theory and was called *Cybernetics*; the latter followed the evolution of digital computer programming (and was symbolic). At that time significant work was being done on both sides of the Atlantic. In the UK a young Christian physicist, Donald M. MacKay at King's College London, was moving from Information Theory and analog computing to look at intelligence and intelligent systems and in the late 1940s developed an early (analog) simple machine learning system. We will look at his contributions in more detail later.[26]

21. Clark , *Thales to Dewey*.
22. Robert Chambers, *Vestiges of the Natural History of Creation and Other Evolutionary Writings* (Chicago, IL: University of Chicago Press, 1994).
23. Lovelace is often described as the world's first programmer, though that is not strictly true. Besides, systems analysis can be seen as a higher-level intellectual activity.
24. Christopher Hollings, Ursula Martin and Adrian Rice, *Ada Lovelace: The Making of a Computer Scientist* (Oxford: Bodleian Library, 2018).
25. Warren S. McCulloch and Walter Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', *Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–33.
26. Donald M. MacKay, *Information, Mechanism and Meaning* (Cambridge, MA: MIT Press, 1969).

In 1950 Alan Turing (1912–1954) published his seminal paper[27] in which he described what has come to be known as the Turing test. In order to pass the Turing test a putatively intelligent machine must operate at a level such that when confronted by a machine and a human interlocutor, a human interrogator is unable to distinguish the two. This, of course, only provides an operational definition of intelligence. If any machine were to actually pass the test, this would not provide conclusive evidence that it was sentient.

One of the earliest successes was the *Logic Theorist* computer program by Allen Newell and Herbert Simon.[28] This theorem proving system could construct proofs of the theorems contained in Bertrand Russell and Alfred Whitehead's *Principia Mathematica*. In fact, several of the proofs were shorter and more elegant than the ones in *Principia Mathematica*.

A major international conference, that attracted delegates from around the world, and set the agenda for AI, took place at Dartmouth College in 1956. It was at this conference that the term 'Artificial Intelligence' as the name for the research area was proposed by John McCarthy (1927–2011) of Stanford University. Some delegates, such as Newell and Simon, preferred the term 'Complex Information Processing' as being less provocative. However, McCarthy (who invented the AI programming language LISP) held the day. Such was the optimism surrounding the early successes of AI in logic and games such as checkers (draughts) that there was a real expectation that the construction of a system that could pass the Turing test was just around the corner, and they predicted that this would happen within ten years! As expected by some, this prediction along with the name did cause consternation in some quarters, leading to IBM sales personnel adopting Peter Drucker's adage: 'A computer is an electronic moron'.

Early success with simple problems[29] is one thing, scaling up to general-purpose real-world scenarios is another. Soon the issue of computational complexity raised its head and progress slowed. But progress there was. One particular example, ELIZA, serves to demonstrate how unexpected reactions can turn someone against AI.

ELIZA[30] was developed by Joseph Weizenbaum (1923–2008) at MIT in the mid-1960s as a natural language processing system to show the straightforward nature of human-computer communication.[31] It was designed to behave like a psychotherapist. However, when people started 'opening up' to ELIZA, it caused some concern, because they continued to do so even after being told that there was no intelligence there. This came to a head when Weizenbaum's secretary asked him to 'close the door' so that

27. Alan Turing, 'Computing Machinery and Intelligence', *Mind* 59.236 (1950), pp. 433–60.
28. Allen Newell and Herbert A. Simon, *The Logic Theory Machine: A Complex Information Processing System*, technical report P-868 (RAND Corporation, June 1956).
29. One may not consider the proofs in *Principia Mathematica* simple, but they were only propositional and to function more generally one needs to be able to work with predicate logic as a minimum.
30. The system was named ELIZA after Eliza Doolittle in *My Fair Lady*, which was a retelling of the story of Pygmalion.
31. Joseph Weizenbaum, 'ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine', *Communications of the ACM* 9.1 (1966), pp. 36–53.

she could talk to ELIZA in private. After this Weizenbaum turned his back on AI and wrote a book warning of its dangers.[32]

The failure of AI systems to live up to the original hype led to an 'AI winter', where funding was hard to come by (and there has been more than one of those). One significant, though unintended, consequence of Marvin Minsky and Seymour Papert's book *Perceptrons*[33] was for funding to dry up almost completely for that research area. Minsky and Papert had shown that at the then current state of development there were several key problems that perceptrons would not be able to solve, in particular those requiring adaption. It was not until 15 years later, when David Rumelhart and James McClelland[34] solved this problem by applying a method that had been devised for assigning credit and blame in economic systems (the *backpropagation algorithm*), that funding started to flow again for Artificial Neural Networks (ANN) research. And since then ANN researchers have not looked back.

On the symbolic side, with the Japanese push in the 'Fifth Generation' project there was a burgeoning of research in what came to be called Expert Systems.[35] These systems used (sophisticated) networks of rules to make the deductions and draw conclusions that an expert in a particular domain might make. These systems achieved reasonable success (around 70–80% on average) in specialist areas and spawned a group of people called 'Knowledge Engineers'. They devised methods to elicit the knowledge that was embedded in the minds of experts. However, this was a tedious and time-consuming task for the simple reason that experts are not good at articulating their expertise. This gave rise to the 'Knowledge Acquisition bottleneck'. It was eventually realised that observing what experts did, and collecting this as data that could be fed into a machine-learning engine, was a more effective means of acquiring the necessary knowledge.

Another issue that arose was that experts could be precious about their expertise especially if they thought they might be outdone by a machine (not that there was much risk of that then… or even now). In order to obviate this issue, Perry Miller proposed a 'Critiquing' approach to medical diagnosis. The system ATTENDING[36] would review a proposed expert diagnosis and suggest modifications or other possibilities where necessary. It also provided a means of explaining the diagnosis. In this way the Expert System was more obviously a tool, like any other, in the medical toolbox rather than a perceived threat.

As things moved on, in the 1990s there was a focus on agency as a key approach to any future AI systems. Agent-based systems have been an active area of research ever

32. Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (New York: W.H. Freeman & Company, 1976).
33. Marvin Minsky and Seymour Papert, *Perceptrons* (Cambridge, MA: MIT Press, 1969). 'Perceptron' is another name for Artificial Neural Nets, though it is less commonly used these days.
34. David E. Rumelhart and James L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, vol. 1 (Cambridge, MA: MIT Press, 1986).
35. Edward A. Feigenbaum and Pamela McCorduck, *The Fifth Generation* (Reading, MA: Addison-Wesley, 1983).
36. Perry L. Miller, 'ATTENDING: Critiquing a Physician's Management Plan', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5.5 (1983), pp. 449–61.

since,[37] including sub-areas such as argumentation.[38] But, whilst these were important technical developments, what has caught the public imagination has been successes in the realm of games: Chess, Jeopardy, and Go. A significant feature of these successes is the fact that both symbolic and neural methods are involved.

Neural Nets were the original Nature Inspired approach. However, since John Holland introduced Genetic Algorithms in the mid-1970s[39] there has been an explosion in the creation of Nature Inspired Algorithms.[40] Just about any natural process is fair game for being turned into an algorithm, for example the immune system, ant colonies, swarms, and even chemical reactions, to name but a few. It has become a bit of a cottage industry.

In the new millennium interest in neural nets grew with the development of Deep Learning (DL).[41] These are basically large-scale neural nets with some sophisticated feedback mechanisms. DL became a practical possibility when Graphical Processing Units (GPUs), which as the name suggests had been developed for high-speed graphical processing, started to be used instead of CPUs as low-cost, high-speed, general-purpose computers. This has proved such a popular area of research that by the mid-2010s there were around 2000 papers per month being published on DL and its applications.[42]

Two key problems with DL are: 1) No one quite knows why it works, and 2) As with all ANN type approaches, the knowledge is embedded in the structure of the net and so has no explicit representation. This means that explanations and justifications are difficult (if not impossible) to obtain. This, as is readily pointed out by those working in symbolic AI, is a major issue, and a reason why DL on its own cannot be the only approach used for real-world AI systems. The fact that DL has been used to interpret radiographs, led one of the early pioneers, Geoff Hinton, to predict (over-optimistically) that soon there would be no need for radiologists.[43] Nonetheless this has had an impact on the profession, with a significant downturn in the number of medics choosing radiology as a career.

37. Russell and Norvig, *Artificial Intelligence*.
38. Iyad Rahwan and Guillermo R. Simari (eds.), *Argumentation in Artificial Intelligence* (London: Springer International Publishing, 2009).
39. John H. Holland, *Adaption in Natural and Artificial Systems* (Cambridge, MA: MIT Press, 1992).
40. Wael Korani and Malek Mouhoub, 'Review on Nature-Inspired Algorithms', *Operations Research Forum* 2.3 (2021), p. 36.
41. Jürgen Schmidhuber, 'Deep Learning in Neural Nets: An Overview', *Neural Nets* 61 (2015), pp. 85–117.
42. M. Mutlu Yapici, Adem Tekerek and Nurettin Topaloglu, 'Literature Review of Deep Learning Research Areas', *Gazi Mühendislik Bilimleri Dergisi* 5.3 (2019), pp. 188–215.
43. 'Let me start by saying a few things that seem obvious. I think if you work as a radiologist, you're like the coyote that's already over the edge of the cliff but hasn't yet looked down, so doesn't know there's no ground underneath him. People should stop training radiologists now. It's just completely obvious that within 5 years, deep learning is going to do better than radiologists, because it's going to be able to get a lot more experience. It might be 10 years, but we've got plenty of radiologists already. I said this at a hospital, and it didn't go down too well.' Geoff Hinton on radiology in 2016, available at: https://www.youtube.com/watch?v=2HMPRXstSvQ.

The Loebner prize was created by Hugh Loebner in 1990 to reward any AI system that could pass the Turing test as adjudicated by a panel of judges. The competition was run annually up until 2020. There were three prizes: bronze for the best AI system in the competition that year; silver for any system that could fool at least two judges; and gold for the first system to pass the Turing test. Only a couple of systems have managed to gain a silver medal, and none has come anywhere close to winning gold. The ones that gained silver did so because, as with ELIZA, there is a tendency in humans to ask the obvious questions that a system developer could prepare for: 'Is there a God?', or 'What are your career aspirations?' These sorts of question do not yield informative answers relative to the task of deciding whether it is human (remember that the Turing test challenge is to decide which, if any, of two interlocutors is human). Luciano Floridi identified that asking questions that would yield maximal information (such as '"The four capitals of the UK are three, Manchester and Liverpool. What's wrong with this sentence?"'[44]) were very effective in quickly revealing which was the AI system. The main take-home lesson from the series of Loebner events is that real Artificial Intelligence is still a long way off, and more likely unachievable.

Successful and Not so Successful Applications

The first successful application of AI was IBM's Deep Blue[45] which was the first system to beat a world chess champion (Garry Kasparov). Buoyed by this success the IBM executives looked for another challenge and decided to focus on Jeopardy. This took seven years. The first attempts were slow, which was no use for a game like Jeopardy where speed is of the essence. Eventually the relevant searches were speeded up sufficiently to enable the Al system, named 'Watson', after the founder of IBM, to take part in the real game. In 2011 Watson became Jeopardy champion: the first AI system to win a game that used natural language to communicate.[46] This caused quite a stir, but it was soon overshadowed by the success of AlphaGo. This was a Deep Learning system developed by DeepMind. It was thought that Go, being a more sophisticated game than chess, would be beyond the capability of any AI system to beat a champion in the near future (if it would ever be achieved[47]).

These examples certainly gave a boost to AI research. However, they were focused on very specific single problems and in that context they were able to do more 'practice' than a human would. For example, Alpha and AlphaGo would play millions of games against another Alpha or AlphaGo system, each learning, by means of reinforcement, the best strategies. This is orders of magnitude more 'experience' than even the best professionals (such as Kasparov) could gain in a lifetime. While it is still a major achievement, that fact puts it into perspective.

44. Luciano Floridi, *The 4th Revolution* (Oxford: Oxford University Press, 2014), pp. 133–34.
45. Feng-Hsiung Hsu, *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion* (Princeton, NJ: Princeton University Press, 2002).
46. David A. Ferrucci, 'Introduction to "This is Watson"', *IBM Journal of Research and Development* 56.3–4 (2012).
47. David Silver, Aja Huang, Chris Maddison et al., 'Mastering the Game of Go with Deep Neural Networks and Tree Search', *Nature* 529.7587 (2016), pp. 484–89.

On the other hand, these headline-making achievements need to be balanced by equally headline-making failures. It is perhaps evidence of hubris on the part of the developers that the following situations occurred.

There was the Amazon recruiting tool, used to select candidates for interview from a set of applicants. It was based on the result of data mining from the features of current Amazon employees. Unfortunately, since most of these employees were male, that came up as the best indicator, hence female applicants were being eliminated before they got to interview.[48] This was an extreme form of bias. However, despite the negative publicity, this is not a problem with AI. Rather, it is an issue with what question was posed and how the system was used. If the same data had been used to answer the question: 'What is the current demographic of Amazon and how can we improve diversity?' the same system could have provided very useful answers.

There was also the Microsoft chatbot, Tay, which had to be withdrawn after a couple of days because it had been turned into a fascist, misogynist chatbot.[49] More recently Meta released Galactica, ostensibly to aid in the writing of scientific papers. Again, its use was quickly suspended after receiving strong criticism from scientists for the sort of 'paper' it was producing.[50]

These are just some of the high-profile issues, but there are other example of real and perceived bias in situations such as mortgage applications This has led to a new, and growing, area of *Accountable and Explainable AI*, with a significant focus on 'fairness'.[51]

Some AI Specialities

In the eighty-odd years since the inception of AI in its modern form, although the initial hopes for general AI have not been realised, it has expanded greatly in terms of the breadth of specialities that now exist with AI as the umbrella term (albeit what would now be called Narrow AI).[52] So much so that many researchers can spend their whole careers focused on one narrowly specified group of problems (with only a cursory knowledge

48. Jeffrey Dastin, '*Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women'*, *Reuters*, October 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed 11 January 2023).
49. Ellie Hunt, '*Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter'*, *The Guardian*, March 2016, https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed 11 January 2023).
50. Matthew Sparkes, '*Meta's Galactica AI Can Write Scientific Papers, But Is It Any Good?'*, *New Scientist*, November 2022, https://www.newscientist.com/article/2347520-metas-galactica-ai-can-write-scientific-papers-but-is-it-any-good/ (accessed 11 January 2023).
51. Adinath Ghadage et al., 'Multi-stage Bias Mitigation for Individual Fairness in Algorithmic Decisions', in N. El Gayar et al., *Artificial Neural Networks in Pattern Recognition. ANNPR 2022*. Lecture Notes in Computer Science, vol. 13739 (2023), pp. 40–52.
52. What has come to be called general AI, or Artificial General Intelligence (AGI) is simply the continuation of the aim of AI in the 1950s: to construct an artificially intelligent being. The lack of success in that enterprise and the obvious success of AI tools in specific areas, such as those listed, led to the adoption of the contrasting term 'Narrow AI' for these.

of the other specialisms). The range of topics studied ranges from the original theorem proving through to Natural Language Generation[53] for persuasion[54] (remember ELIZA), and more recently explorations of Artificial Emotion.[55] The first of these[56] has been applied to Natural Theology in the form of a non-modal automated version of the ontological argument.[57] In the world of Machine Learning there has been significant work done in symbolic learning applied to the world of scientific theory formation.[58] Now, one of the Turing Institute's Grand Challenges is to have a robot scientist produce Nobel Prize worthy research by 2050. Another Grand Challenge is in the area of Foundational Models and Large Language Models. The latter is creating a fair amount of chatter on both social and mainstream media with the release of chat-GPT, not least in the academic world where there are renewed concerns about monitoring and marking student essays.[59]

Donald M. MacKay

One key Christian contributor to the development of AI was Donald M. MacKay (1922–1987). He studied Physics and Electronics at St Andrew's University followed by a R&D post at the Admiralty Research Establishment working on Radar during World War II. It was during this time that he formed an interest in Information Theory, though he later expanded his interests to explore the question of mind-like behaviour in artefacts and was one of the UK representatives at the Dartmouth conference. During this time he developed the ideas that he would defend for the rest of his life: that Energy and Information were complementary, and that because of this even if determinism were in fact the case, agents would still be free (where 'agents' here could be either human or artificial). Finally he shifted his interest to brain science, and had an eminent career applying the method of Information Theory to the study of the brain.

53. Ehud Reiter and Robert Dale, 'Building Applied Natural Language Generation Systems', *Natural Language Engineering* 3.1 (1997), p. 57.
54. Ehud Reiter, Roma Robertson and Liesl Osman, 'Types of Knowledge Required to Personalise Smoking Cessation Letters', *Artificial Intelligence in Medicine* 16 (1999), pp. 389–99.
55. Rosalind W. Picard, 'Toward Machines with Emotional Intelligence', ch. 16 in Gerald Matthews et al. (eds.), *The Science of Emotional Intelligence: Knowns and Unknowns* (Oxford: Oxford University Press, 2004), pp. 396–416.
56. Paul E. Oppenheimer and Edward N. Zalta, 'A Computationally-Discovered Simplification of the Ontological Argument', *Australasian Journal of Philosophy* 89.2 (2010), pp. 333–49.
57. This could be seen as a step towards Llull's aim of automating some theological arguments, albeit not being uniformly deemed successful. See Ted Parent, 'On the PROVER9 Ontological Argument', *Philosophia* 43.2 (2015), pp. 475–83.
58. Kevin Williams, Elizabeth Bilsland and Andrew Sparkes et al., 'Cheaper Faster Drug Development Validated by the Repositioning of Drugs against Neglected Tropical Diseases', *Journal of the Royal Society. Interface* 12.104 (2015); George M. Coghill, Ashwin Srinivasan and Ross D. King, 'Qualitative System Identification from Imperfect Data', *Journal of Artificial Intelligence Research* 32 (2008), pp. 825–77.
59. Maya Yang, '*New York City Schools Ban AI Chatbot that Writes Essays and Answers Prompts'*, *The Guardian*, 6 January 2023, https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt (accessed 11 January 2023).

Complementarity

MacKay's basic idea of Complementarity is that two descriptions are complementary 'if they refer to the same object, each is in principle exhaustive, yet they make different assertions because the context of the concepts used are mutually exclusive, so that significant aspects referred to in one are necessarily omitted from the other'.[60] MacKay would often illustrate this by reference to a neon sign displaying a message, which, for example, might be something like: 'Bloggs coffee is best'. In order to be visible, several physical and chemical features need to be operational so that the neon will glow.[61] An engineer could give a complete, exhaustive, description of the operation of the system in terms of its chemistry and physics without making any reference to the message the sign conveys, because to a large extent that is irrelevant to the physical operation. Conversely, the information in the sign, while embodied in the physical features, is independent of them for meaning; and a linguist for example could give an exhaustive analysis of the meaning of the statement without any reference to the physical aspects of the message.

There is a view, which is still prevalent, that human beings can be accounted for purely as material objects: that the mind is simply (or nothing but) the operations of the brain. This position MacKay referred to as 'nothing buttery', and saw that it is clearly undermined by complementarity. He then considered that minds and brains are complementary aspects of the one entity, yielding and embodiment of the mind. It is not brains that think, but persons! In assessing the relation between mind and body, complementarity means that the physical and mental aspects of humans are correlated but not identical. In this regard he did not consider dualism as incoherent, just unnecessary.[62] This he referred to as 'duality without dualism' and related it to the unity of man as described in Genesis 1 and 2.[63] As such there is a hierarchy in the complementary relations: the thinking person is higher than the brain activity.[64]

60. Donald M. MacKay, 'Complementarity II', *Proceedings of the Aristotelian Society* suppl. vol XXXII (1958), pp. 105–122.
61. Donald M. MacKay, *The Clockwork Image* (London: InterVarsity Press, 1974).
62. Donald M. MacKay, *Behind the Eye* (Oxford: Blackwell, 1991), p. 274, n. 1. In this respect MacKay highlights the importance of the body as part of being human (and the image of God). This is somewhat in contrast to some recent theological speculations on 'superintelligence', e.g. Calvin Mercer and Tracy Trothen, *Religion and the Technological Future* (London: Palgrave Macmillan, 2021) and Yong Sup Song, 'Religious AI and Option to the Risks of Superintelligence', *Theology and Science* 19.1 (2021), pp. 65–78 which focus only on the putative mental aspects.
63. Donald M. MacKay, 'The Sovereignty of God in the Natural World', *Scottish Journal of Theology* 21.1 (1968), pp. 13–26.
64. This led to accusations that MacKay was himself a materialist (or semi-materialist), e.g. William A. Dembski, 'Converting Matter into Mind: Alchemy and the Philosopher's Stone in Cognitive Science', *Perspectives on Science and Christian Faith* 42.4 (1990), pp. 202–226. However, he also viewed the soul as being embodied in the mind, which is not commensurate with materialism; see Donald M. MacKay, *The Open Mind and Other Essays: A Scientist in God's World*, ed. Melvin Tinker (Leicester: InterVarsity Press, 1988), p. 71.

Mackay did not consider AI to be problematic in principle. He did not consider it to be in any way encroaching on God's prerogative in creation; rather he considered it a form of *procreation*, which is something humans are rather good at. He was, however, careful to distinguish between intelligent agency and imitation. A favourite question was 'In Hamlet's soliloquy, how many agents are on the stage?' The answer, of course, is 'One'. There is no personal Hamlet, merely an actor imitating him.[65] Mackay's considered view on the matter may be summarised as: There is no theological ground to deny 'artificial begetting' as a possibility. But in the current state of play we are a long way from this. Nonetheless, semantic information theory and neuroscience can suggest what is required for an artefact to be considered to embody a conscious agent. However, 'Whether these requirements can be met in anything other than biological material is an open question'.[66]

Logical Indeterminism

MacKay devised a thought experiment in which he considered a scenario where one might try to get a picture of a person's brain by means of what he called a *cerebroscope*.[67] From his deliberations on this thought experiment he concluded that it would be possible for an outside observer, a *super-scientist* say, to gain a complete picture of the state of the brain, but, paradoxically, it would not be possible for the person themselves to utilise the cerebroscope and get such a picture (see Figure 1). This is because one's brain cannot be in a particular state at a particular time and be simultaneously observed by oneself to be in that state: in trying to observe it one would continuously change it. An illustration of this sort of thing is placing a microphone in front of the speaker to which it is connected: a squeal of increasing volume is heard. No stable state can be reached. As MacKay put it: ' there does not exist a complete specification of a person's brain state that they would be correct to believe and incorrect to disbelieve'.[68]

MacKay reflected on whether it really was the case that if the world was fully deterministic this would mean that a person was not free and responsible. His conclusion that it did not mean that followed from his cerebroscope experiment. In a fully deterministic world the super-scientist mentioned above can observe every feature of a person's brain, and correctly predict what that person was going to do in the immediate future. He then asked the question: 'Would the prediction be inevitable?' And his answer was 'No!' The reason is that while the prediction has a claim to the assent of nearly everyone there is one person for whom it is not true, and that is the person who is being observed. If that person were told the prediction, it would immediately become out of date because communicating the prediction would change the brain state that was the basis for the prediction. For them the future remains open until they make the decision (arguably, as 'open' as any libertarian could wish). Here again, whether a particular proposition is

65. MacKay, *The Open Mind and Other Essays*, p. 125.
66. MacKay , *The Open Mind and Other Essays*, p. 136.
67. Donald M. MacKay, 'The Logical Indeterminateness of Human Choices', *British Journal for the Philosophy of Science* 24.4 (1973), pp. 405–408.
68. MacKay , 'The Logical Indeterminateness of Human Choices', p. 405.

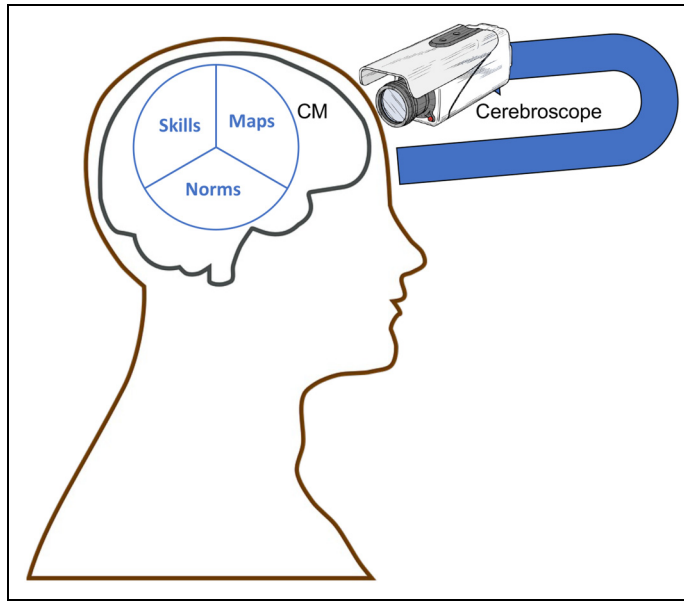


Figure 1. A human agent.

true for a person depends on the standpoint of that person. So, the claim that one couldn't help it because it was determined is false. MacKay also saw this *openness* of the future as having theological implications for the debate between Calvinism and Arminianism.[69]

In like manner, if a truly artificially intelligent agent were to be begotten and the tried to view their own `brain' by means of the cerebroscope, the same rules would apply to them. In fact, even before the stage of sentience, if the AI had the capability to reflect on its own knowledge state, the same paradoxical situation would apply, and such an agent could be considered as free (in MacKay's sense at least).

Recent Fears and Concerns

There has been a resurgence of concern and prediction of dystopian factors of various forms associated with the so-called *singularity*. Several high-profile people including Stephen Hawking and Elon Musk as well as several AI experts wrote an open letter expressing serious concerns about the future of AI, including speculation regarding the 'singularity'.[70] While one may dismiss the concerns of Hawking and Musk because,

69. 'Thus the divine "foreknowledge" of our future, oddly enough, has no unconditional logical claim upon us, unknown to us. This, I believe, demonstrates a fallacy underlying both the theological dispute between Arminianism and Calvinism, and the philosophical dispute between physical or psychological determinism … and libertarianism …' MacKay, *The Clockwork Image*, p. 82.
70. The point at which sentience and superintelligence is achieved.

for all their eloquence and standing in the science or business communities, they are not expert in, nor practitioners in AI, the same cannot be said of another signatory of the letter: Stuart Russell. His position in the AI world is of the highest calibre: Professor of AI at Stanford, major AI researcher and co-author of the leading AI textbook.[71] As noted previously, being a leader in AI does not prevent one from turning against it (e.g., Weizenbaum); however, Russell's argument, stated in his book *Human Compatible,*[72] is fallacious. He correctly points out that within weeks of Ernest Rutherford's claim that there was no practical use from the splitting of the atom, Leo Szilard showed that it could be used to release large amounts of energy, which formed the foundation for the development of both nuclear power and the atomic bomb. Russell then uses this historical example to warn against claims that the AI singularity is nowhere in sight. Of course, there is a small possibility that he may be right, but the main reason that the claim is made is that there is no actual evidence to support the contention of any immanent singularity. Russell is well-versed in logic and an expert in Bayesian reasoning so it is puzzling as to why he would say this.

If one does want to draw parallels with physics this can more appropriately be done by contrasting physics and AI as a whole with certain specific problems within those domains. Russell's example is more akin to the claim, made not that long before AlphaGo beat Lee Sedol, that although Deep Blue beat Kasparov, it would not be possible to win at Go. And science is replete with such examples (e.g., the impossibility of heavier-than-air flight). On the other hand, the general trend in Physics (or Science) as a whole is in the opposite direction. For example, in the nineteenth century there was the real expectation that with a few more *i*'s dotted and *t*'s crossed, Newtonian mechanics would have all the fundamental problems of physics solved. However, while it was not clear at the time, James Clerk Maxwell's electromagnetic theories undermined this expectation and led directly to Einstein's theories of Relativity.

In the twentieth century the quests for the *General Unification Theory* or *Theory of Everything* were seen as real possibilities, but now with the discovery of Dark Energy and Matter we realise that we only have an understanding of around 5 per cent of what is in the universe. It is worse with AI: its much shorter history has been peppered with grandiose claims regarding what lay 'just around the corner', from the 10-year plan of the Dartmouth conference, or the Fifth Generation project, to the current obsession with 'super-intelligence' and the 'singularity'. The only thing that can be said with any certainty is that at the current time the best one can see is that, as MacKay has argued and as is succinctly expressed in the words of Luciano Floridi: 'AI is not logically impossible but it is highly implausible'.[73] This being the case, it seems more appropriate for theologians to focus their attention more on the implications of the current state-of-the-art in AI, such as is being done by, for example, Ximian Xu in exploring the potential of AI

71. Russell and Norvig, *Artificial Intelligence*.
72. Stuart Russell, *Human Compatible: AI and the Problem of Control* (London: Penguin, 2019).
73. Luciano Floridi, 'Should We Be Afraid of AI?', *Aeon Essays,* 9 May 2016, https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible (accessed 23 March 2023).

bots to provide initial pastoral support,[74] rather than on the science-fiction aspects (fun though that may be). There should also be an effort made to understand that current AI research is not homogeneous,[75] and seek to bring theological insights to bear on the whole enterprise. But I leave it to the theological experts to decide how best to engage with that. In the end our attitude should be thankfulness to God for the good that AI can bring, mixed with the cautious stewardship that must direct our use of any technological development.

ORCID iD

George M. Coghill https://orcid.org/0000-0002-2047-8277

74. Ximian Xu, 'A Theological Account of Artificial Moral Agency', *Studies in Christian Ethics*, 2023, https://doi.org/10.1177/09539468231163002.
75. See section: 'What is Artificial Intelligence?' above.