Artificial Intelligence is stupid and causal reasoning wont fix it

J. Mark Bishop[1]

[1] TCIDA, Goldsmiths, University of London, London, UK
Email: m.bishop@gold.ac.uk

Abstract

Artificial Neural Networks have reached 'Grandmaster' and even 'super-human' performance' across a variety of games, from those involving perfect-information, such as Go [*Silver et al.*, 2016]; to those involving imperfect-information, such as 'Starcraft' [*Vinyals et al.*, 2019]. Such technological developments from AI-labs have ushered concomitant applications across the world of business, where an 'AI' brand-tag is fast becoming ubiquitous. A corollary of such widespread commercial deployment is that when AI gets things wrong - an autonomous vehicle crashes; a chatbot exhibits 'racist' behaviour; automated credit-scoring processes 'discriminate' on gender etc. - there are often significant financial, legal and brand consequences, and the incident becomes major news.

As Judea Pearl sees it, the underlying reason for such mistakes is that "*... all the impressive achievements of deep learning amount to just curve fitting*". The key, Pearl suggests [*Pearl and Mackenzie*, 2018], is to replace 'reasoning by association' with 'causal reasoning' - the ability to infer causes from observed phenomena. It is a point that was echoed by Gary Marcus and Ernest Davis in a recent piece for the New York Times: "*we need to stop building computer systems that merely get better and better at detecting statistical patterns in data sets – often using an approach known as "Deep Learning" – and start building computer systems that from the moment of their assembly innately grasp three basic concepts: time, space and* ***causality***" [*Marcus and Davis*, 2019].

In this paper, foregrounding what in 1949 Gilbert Ryle termed 'a category mistake' [*Ryle*, 1949], I will offer an alternative explanation for AI errors; it is not so much that AI machinery cannot 'grasp' causality, but that AI machinery (qua computation) cannot understand anything at all.

Keywords: Cognitive Science, Artificial Intelligence, Artificial Neural Networks, Causal Cognition, Chinese Room Argument, Dancing with Pixies.

1 Making a mind

For much of the twentieth century the dominant cognitive paradigm identified the mind with the brain; as the Nobel laureate Francis Crick eloquently summarised:

> "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules. As Lewis Carroll's Alice might have phrased, "You're nothing but a pack of neurons". This hypothesis is so alien to the ideas of most people today that it can truly be called astonishing" *Crick* [1994].

Motivation for the belief that a computational simulation of the mind is possible stemmed initially from the work of *Turing* [1937] and *Church* [1936] and the 'Church-Turing hypothesis'; in Turing's formulation, every 'function which would naturally be regarded as computable' can be computed by the 'Universal Turing Machine'. If computers can adequately model the brain then, theory goes, it ought to be possible to *program* them to act like minds. And consequently, in the latter part of the twentieth century, Crick's "Astonishing Hypothesis" helped fuel an explosion of interest in connectionism: both high-fidelity simulations of the brain (computational neuroscience; theoretical neurobiology) and looser - merely 'neural inspired' - analogues (cf. Artificial Neural Networks; Multi-Layer Perceptrons; 'Deep Learning' systems).

But the fundamental question that Crick's hypothesis raises is, of course, this: if we ever succeed in fully instantiating a *sufficiently accurate* simulation of the brain on a digital computer, will we also have fully instantiated a digital [computational] mind, with all the human mind's causal power of teleology, understanding and reasoning; will AI finally have succeeded in delivering 'Strong AI'[1]

Of course, *if* Strong AI is possible, accelerating progress in its underpinning technologies[2] - entailed both by the use of AI systems to design ever more sophisticated AIs and the continued doubling of raw computational power every two years[3] - will eventually cause a runaway effect whereby the

[1] Strong AI, a term coined by *Searle* [1980] in the 'Chinese Room Argument' (CRA), entails that, "*... the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states*", which Searle contrasted with "Weak AI" wherein "*... the principal value of the computer in the study of the mind is that it gives us a very powerful tool.*" I.e. Weak AI focusses on epistemic issues relating to engineering a simulation of [human] intelligent behaviour whereas Strong AI, in seeking to engineer a computational system with all the causal power of a mind, focusses on the ontological.

[2] Cf. "[A]mplifiers for intelligence - devices that supplied with a little intelligence will emit a lot", [*Ashby*, 1956].

[3] Cf. Moore's 'law' :- the observation that the number of transistors in a dense integrated circuit approximately doubles every two years.

Artificial Intelligence will inexorably come to exceed human performance on all tasks[4]; the so-called point of [technological] 'singularity' ([in]famously predicted by Ray Kurzweil to occur as soon as 2045[5]). And, at the point this 'singularity' occurs, so commentators like Kevin Warwick[6] and Stephen Hawking[7] suggest, humanity will, effectively, have been "superseded" on the evolutionary ladder and be obliged to eke out its autumn days listening to 'Industrial Metal' music and gardening; or, in some of Hollywood's even more dystopian dreams, cruelly subjugated (and/or exterminated) by 'Terminator' machines.

In this paper, however, I will offer a few 'critical reflections' on one of the central, albeit awkward, questions of AI: why is it that, over seven decades since Alan Turing first deployed an 'effective method' to play chess in 1948, we have seen enormous strides in engineering particular machines to do clever things – from driving a car to beating the best at Go – but almost no progress in getting machines to genuinely understand; to seamlessly apply knowledge from one domain into another – the so-called problem of 'Artificial General Intelligence' (AGI); the skills that both Hollywood and the wider media really think of, and depict, as Artificial Intelligence?

2 Neural Computing

The earliest Cybernetic work in the burgeoning field of 'neural computing' lay in various attempts to understand, model and emulate neurological function and learning in animal brains, the foundations of which were laid in 1943 by the neurophysiologist Warren McCulloch and the mathematician

[4]Conversely, as Francois Chollet, a senior engineer at Google and well known sceptic of the 'Intelligence Explosion' scenario; trenchantly observed in 2017: "*The thing with recursive self-improvement in AI, is that if it were going to happen, it would already be happening. I.e. Auto Machine Learning systems would come up with increasingly better Auto Machine Learning systems, Genetic Programming would discover increasingly refined GP algorithms*" and yet, as Chollet insists, "*no human, nor any intelligent entity that we know of, has ever designed anything smarter than itself*".

[5]*Kurzweil* [2005] "set the date for the Singularity - *representing a profound and disruptive transformation in human capability* - as 2045".

[6]In his 1997 book "March of the Machines" [*Warwick*, 1997] observed that there were already robots with the '*brain power of an insect*'; soon, or so he predicted, there would be robots with the '*brain power of a cat*', and soon after that there would be '*machines as intelligent as humans*'. When this happens, Warwick darkly forewarned, the science-fiction nightmare of a 'Terminator' machine could quickly become reality because such robots will rapidly, and inevitably, become more intelligent and superior in their practical skills than the humans who designed and constructed them.

[7]In a television interview with Professor Stephen Hawking on December 2nd 2014, Rory Cellan-Jones asked how far engineers had come along the path towards creating Artificial Intelligence, to which Professor Hawking alarmingly replied, "*Once humans develop artificial intelligence it would take off on its own and redesign itself at an ever increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded.*"

Walter Pitts [*McCulloch and Pitts*, 1943].

Neural Computing defines a mode of problem solving based on 'learning from experience' as opposed to classical, syntactically specified, 'algorithmic' methods; at its core is "*the study of networks of 'adaptable nodes' which, through a process of learning from task examples, store experiential knowledge and make it available for use*" [*Aleksander and Morton*, 1995]. So construed, an 'Artificial Neural Network' (ANN) is constructed merely by appropriately connecting a group of adaptable nodes ('artificial neurons').

- A *single layer neural network* only has one layer of adaptable nodes between the input vector, X and the output vector O, such that the output of each of the adaptable nodes defines one element of the network output vector O.

- A *multi-layer neural network* has one or more 'hidden layers' of adaptable nodes between the input vector and the network output; in each of the network *hidden layers*, the outputs of the adaptable nodes connect to one or more inputs of the nodes in subsequent layers and in the network *output layer*, the output of each of the adaptable nodes defines one element of the network output vector O.

- A *recurrent neural network* is a network where the output of one or more nodes is fed-back to the input of other nodes in the architecture, such that the connections between nodes form a 'directed graph along a temporal sequence', so enabling a recurrent network to exhibit 'temporal dynamics'; enabling a recurrent network to be sensitive to particular *sequences* of input vectors.

Since 1943 a variety of frameworks for the adaptable nodes have been proposed[8] however the most common, as deployed in many 'deep' neural networks, remain grounded on the McCulloch/Pitts model.

2.1 The McCulloch/Pitts (MCP) model

In order to describe how the basic processing elements of the brain might function, McCulloch and Pitts showed how simple electrical circuits, connecting groups of 'linear threshold functions', could compute a variety of logical functions [*McCulloch and Pitts*, 1943]. In their model McCulloch

[8] These include: 'spiking neurons' as widely used in computational neuroscience [*Hodgkin and Huxley*, 1952]; 'kernel functions' as deployed in 'Radial Basis Function' networks [*Broomhead and Lowe*, 1988] and 'Support Vector Machines' [*Boser et al.*, 1992]; 'Gated MCP Cells', as deployed in LSTM networks [*Hochreiter and Schmidhuber*, 1997]; 'n-tuple' or 'RAM' neurons, as used in 'Weightless' neural network architectures [*Bledsoe and Browning*, 1959; *Aleksander and Stonham*, 1979] and 'Stochastic Diffusion Processes' [*Bishop*, 1989] as deployed in the NESTOR multi-variate connectionist framework [*Nasuto et al.*, 2009].

and Pitts provided a first (albeit very simplified) mathematical account of the chemical processes that define neuronal operation and in so doing realised that the mathematics that describe the neuron operation exhibited exactly the same type of logic that Shannon deployed in describing the behaviour of switching circuits: namely, the calculus of propositions.

McCulloch and Pitts realized (*ibid*) (a) that neurons can receive positive or negative encouragement to fire, contingent upon the type of their 'synaptic connections' (excitatory or inhibitory) and (b) that in firing the neuron has effectively performed a 'computation'; once the effect of the excitatory/inhibitory synapses are taken into account, it is possible to *arithmetically* determine the net effect of incoming patterns of 'signals' innervating each neuron.

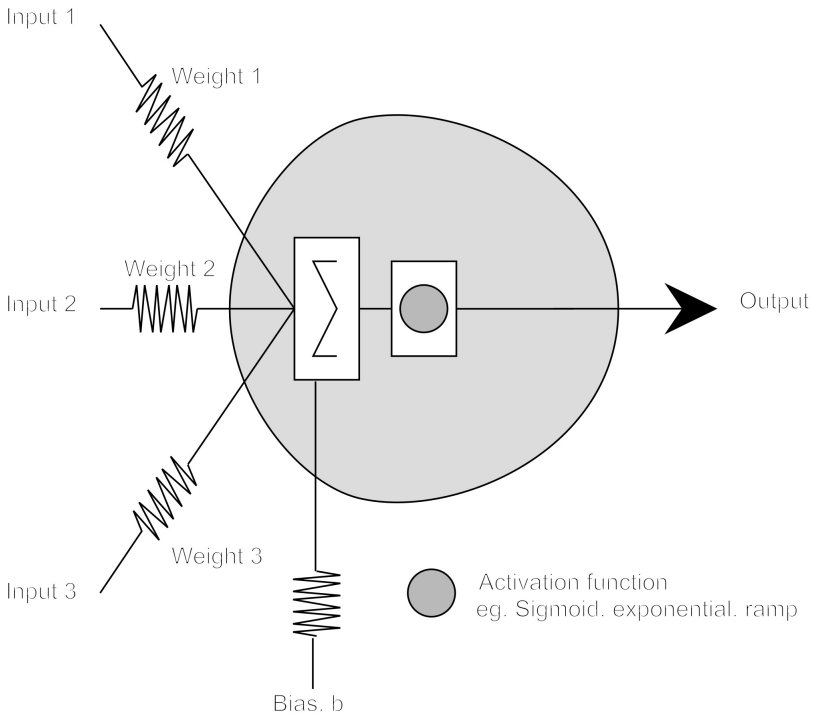

Figure 1: The McCulloch-Pitts neuron model.

In a simple MCP threshold model, adaptability comes from representing each synaptic junction by a variable (usually rational) valued weight W_i, indicating the degree to which the neuron should react to the $_i th$ particular input (see Figure 1). By convention, positive weights represent excitatory synapses and negative, inhibitory synapses; the neuron firing threshold being represented by a variable T. In modern use T is usually clamped to zero and a threshold implemented using a variable 'bias' weight, b; typically, a

neuron firing[9] is represented by the value +1 and not firing by 0.

Activity at the i_{th} input to an n input neuron is represented by the symbol X_i and the effect of the i_{th} synapse by a weight W_i, hence the net effect of the i_{th} input on the i_{th} synapse on the MCP cell is thus $X_i \times W_i$. Thus the MCP cell is denoted as firing if:

$$\sum_{i}^{n} X_i \times W_i + b \quad \geq \quad 0 \tag{1}$$

In a subsequent generalisation of the basic MCP neuron, cell output is defined by a further [typically non-linear] function of the weighted sum of its input; the neuron's *activation function*.

McCulloch and Pitts proved (*ibid*) that if 'synapse polarity' is chosen appropriately, any single pattern of input can be 'recognised' by a suitable network of MCP neurons (i.e. any finite logical expression can be realised by a suitable network of McCulloch-Pitts neurons). In other words, the McCulloch-Pitts' result demonstrated that networks of artificial neurons could be mathematically specified which would perform 'computations' of immense complexity and power and in so doing, opened the door to a form of problem solving based on the design of appropriate neural network architectures and automatic (machine) 'learning' of appropriate network parameters.

3 Embeddings in Euclidean space

The most commonly used framework for information representation and processing in artificial neural networks (via generalised McCulloch/Pitts neurons) is a subspace of Euclidean space. Supervised learning in this framework is equivalent to deriving appropriate transformations (learning appropriate mappings) from training data (problem exemplars; pairs of $Input + `Target\,Output'$ vectors). The majority of learning algorithms adjust neuron interconnection weights according to a specified 'learning rule', the adjustment in a given time step being a function of a particular training example.

Weight updates are successively aggregated in this manner until the network reaches an equilibrium, at which point no further adjustments are made or, alternatively, learning stops before equilibrium to avoid 'overfitting' the training data. On completion of these computations, knowledge about the training set is represented across a distribution of final weight values; thus, a trained network does not possess any internal representation of the (potentially complex) relationships *between* particular training exemplars.

[9] "In psychology .. the fundamental relations are those of two valued logic" and McCulloch and Pitts recognised neuronal firing as equivalent to 'representing' a proposition as $TRUE$ or $FALSE$ [*McCulloch and Pitts*, 1943].

Classical multi-layer neural networks are capable of discovering non-linear, continuous transformations between objects or events, but nevertheless they are restricted by operating on representations embedded in the linear, continuous structure of Euclidean space. It is, however, doubtful whether regression constitutes a satisfactory (or the most general) model of information processing in natural systems.

As Nasuto et al. observed *Nasuto et al.* [1998], the world, and relationships between objects in it, is fundamentally non-linear; relationships between real-world objects (or events) are typically far too messy and complex for representations in Euclidean spaces - and smooth mappings between them - to be appropriate embeddings (e.g. entities and objects in the real-world are often fundamentally discrete or qualitatively vague in nature, in which case Euclidean space does not offer an appropriate embedding for their representation).

Furthermore, representing objects in a Euclidean space imposes a serious additional effect, because Euclidean vectors can be compared to each other by means of *metrics*; enabling data to be compared in spite of any real-life constraints (sensu stricto, metric rankings may be undefined for objects and relations of the real-world). I.e. As Nasuto et al. highlight (*ibid*), it is not usually the case that all objects in the world can be equipped with a 'natural ordering relation'; after all, what is the natural ordering of 'banana' and 'door'?

It thus follows that classical neural networks are best equipped only for tasks in which they process numerical data whose relationships can be reflected by Euclidean distance. In other words, classical connectionism can be reasonably-well applied to the same category of problems which could be dealt with by various regression methods from statistics; as Francois Chollet[10], in reflecting on the limitations of deep learning, recently remarked:

> "[a] deep learning model is 'just' a chain of simple, continuous geometric transformations mapping one vector space into another. All it can do is map one data manifold X into another manifold Y, assuming the existence of a learnable continuous transform from X to Y, and the availability of a dense sampling of X: Y to use as training data. So even though a deep learning model can be interpreted as a kind of program, inversely most programs cannot be expressed as deep learning models-for most tasks, either there exists no corresponding practically-sized deep neural network that solves the task, or even if there exists one, it may not be learnable ... most of the programs that one may wish to learn cannot be expressed as a continuous geometric morphing of a data manifold." [*Chollet*, 2018].

Over the last decade, however, Artificial Neural Network technology has developed beyond performing 'simple function approximation' (cf. Multi-

[10]Chollet is a senior software engineer at Google, who - as the primary author and maintainer of Keras, the Python open source neural network interface designed to facilitate fast experimentation with Deep Neural Networks - is particularly familiar with the problem-solving capabilities of Deep Learning systems.

Figure 2: Terrence Broad's Auto-encoding network 'dreams' of Bladerunner (from [*Broad*, 2016]).

Layer Perceptrons) and deep [discriminative[11]] classification (cf. Deep Convolutional Networks), to include new, *Generative* architectures[12] where - *because they can learn to generate any distribution of data* - the variety of potential use-cases is huge (e.g. generative networks can be taught to create novel outputs similar to real-world exemplars across any modality: images, music, speech, prose etc).

3.1 Autoencoders, Variational Autoencoders and Generative Adversarial Networks

On the right hand side of Figure (2) we see the output of a neural system, engineered by Terence Broad whilst studying for a MSc at Goldsmiths. Broad used a 'complex, deep auto-encoder neural network' to process Blade Runner - a well-known sci-fi film which riffs on the notion of what is human and what is machine - building up its own 'internal representations' of that film and then re-rendering these to produce an output movie that is surprisingly similar to the original (shown on the left).

In his dissertation *Broad* [2016], a 'Generative Autoencoder Network' reduced each frame of Ridley Scott's Blade Runner to 200 'latent variables' (hidden representations), then invoked a 'decoder network' to reconstruct each frame just using those numbers. The result is eerily suggestive of an

[11] A discriminative architecture - or discriminative classifier without a model - can be used to "discriminate" the value of the target variable Y, given an observation x.

[12] A generative architecture can be used to "generate" random instances, either of an observation and target (x, y), or of an observation x given a target value y.

Android's dream; the network, working without human instruction, was able to capture the most important elements of each frame so well that when its reconstruction of a clip from the Blade Runner movie was posted to Vimeo, it triggered a 'Copyright Takedown Notice' from Warner Brothers.

To understand if Generative Architectures are subject to the Euclidean constraints identified above for classical neural paradigms, it is necessary to trace their evolution: from the basic Autoencoder Network, through Variational Autoencoders to Generative Adversarial Networks.

3.1.1 Autoencoder Networks

'Autoencoder Networks' [*Kramer*, 1991] create a latent (or hidden), typically much compressed, representation of their input data. When Autoencoders are paired with a decoder-network, the system can reverse this process and reconstruct the input data that generates a particular latent representation. In operation, the Autoencoder Network is given a data input x, which it maps to a latent representation z, from which the decoder network reconstructs the data input x' (typically, the cost function used to train the network is defined as the mean squared error between the input x and the reconstruction x'). Historically, Autoencoders have been used for 'feature learning' and 'reducing the dimensionality of data' [*Hinton and Salakhutdinov*, 2006], but more recent variants (described below) have been powerfully deployed to learn 'Generative Models' of data.

3.1.2 Variational Autoencoder Networks

In taking a 'variational Bayesian' approach to learning the hidden representation, 'Variational Autoencoder Networks' [*Kingma and Welling*, 2013] add an additional constraint; placing a strict assumption on the distribution of the latent variables. Variational Autoencoder Networks are capable of both compressing data instances (like an Autoencoder) and generating new data instances.

3.1.3 Generative Adversarial Networks

Generative Adversarial Networks [*Goodfellow et al.*, 2014] deploy two 'adversary' neural networks: one - the Generator - synthesises new data instances, whilst the other - the Discriminator - rates each instance as how likely it is to belong to the training dataset. Colloquially, the Generator takes the role of a 'counterfeiter' and the Discriminator the role of 'the police', in a complex and evolving game of cat and mouse, wherein the counterfeiter is evolving to produce better and better counterfeit money while the police are getting better and better at detecting it. This game goes on until, at convergence, both networks have become very good at their tasks; so good that Yann LeCun, Facebook's AI Director of Research, recently claimed them to

What Machine Learning Can Do

A simple way to think about supervised learning.

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

SOURCE ANDREW NG © HBR.ORG

Figure 3: The tasks ANNs and ML can perform.

be "*the most interesting idea in the last ten years in Machine Learning*[13]".

Nonetheless, as Goodfellow emphasizes (*ibid*), the generative modelling framework is most straightforwardly realised using "multilayer perceptron models". Hence, although the functionally of generative architectures moves beyond the simple function-approximation and discriminative-classification abilities of classical multi-layer perceptrons, at heart, in common with all neural networks that learn, and operate on, functions embedded in Euclidean space[14], they remain subject to the constraints of Euclidean embeddings highlighted above.

4 Problem solving using Artificial Neural Networks

In analysing what problems neural networks and machine learning *can* solve, Andrew Ng[15] suggested that if a task only takes a few seconds of human judgement and, at its core, merely involves an association of A with B, then it may well be ripe for imminent AI automation (see Figure (3)).

However, although we can see how we might deploy a trained neural network in the engineering of solutions to specific, well-defined problems - such as, "*Does a given image contain a representation of a human face?*"

[13] Quora July 28, 2016, (`https://www.quora.com/session/Yann-LeCun/1`).

[14] Including neural networks constructed using alternative 'adaptable node' frameworks (e.g. those highlighted in footnote [8]), where these operate on data embeddings in Euclidean space.

[15] Adjunct professor at Stanford University and formerly associate professor and Director of its AI Lab.

- it remains unproven if (a) every human intellectual skill is computable in this way and, if so, (b) is it possible to engineer an *Artificial General Intelligence* that would negate the need to engineer bespoke solutions for each and every problem.

For example, to master image recognition, an ANN might be taught using images from ImageNet (a database of more than 14 million photographs of objects that have been categorised and labelled by humans), but is this how humans learn? In [*Savage*, 2019] Tomaso Poggio, a computational neuroscientist at the Massachusetts Institute of Technology, observes that, although a baby may see around a billion images in the first two years of life, only a tiny proportion of objects in the images will be actively pointed out, named and labelled.

4.1 On cats, classifiers and grandmothers

In 2012, organisers of 'The Singularity Summit', an event which foregrounds predictions from the like of Kurzweil and Warwick (vis a vis 'the forthcoming Technological Singularity' [sic]), invited Peter Norvig[16] to discuss a surprising result from a Google team that appeared to indicate significant progress towards the goal of unsupervised category learning in machine vision; instead of having to engineer a system to recognise each and every category of interest (e.g. to detect if an image depicts a human face, a horse, a car etc.) by training it with explicitly labelled examples of each class (so called, 'supervised learning'), Le et al. conjectured that it might be possible to build high-level image classifiers *using only un-labelled images*, "*... we would like to understand if it is possible to build a face detector from only un-labelled images. This approach is inspired by the neuro-scientific conjecture that there exist highly class-specific neurons in the human brain, generally and informally known as 'grandmother neurons*".

In his address, [*Norvig*, 2012] described what happened when Google's 'Deep Brain' system was 'let loose' on unlabelled images obtained from the internet:

> ".. and so this is what we did. We said we're going to train this, we're going to give our system ten million YouTube videos, but for the first experiment, we'll just pick out one frame from each video. And, you sorta know what YouTube looks like .. We're going to feed in all those images and then we're going to ask it to represent the world. So what happened? Well, this is YouTube, so there will be cats.
>
> And what I have here is a representation of two of the top level features [see Figures (4) and (5)]. So the images come in, they're compressed there, we build up representations of what's in all the images. And then at the top

[16]Peter is Director of Research at Google and, even though also serving an adviser to 'The Singularity University', clearly has reservations about the notion: "*.. this idea, that intelligence is the one thing that amplifies itself indefinitely, I guess, is what I'm resistant to ..*" [Guardian 23/11/12].

Figure 4: Reconstructed archetypal cat (extracted from YouTube video of Peter Norvig's address to the 2012 Singularity summit).

Figure 5: Reconstructed archetypal face (extracted from YouTube video of Peter Norvig's address to the 2012 Singularity summit).

> level, some representations come out. These are basis functions - features that are representing the world - and the one on the left here is sensitive to cats. So these are the images that most excited that this node in the network; that 'best matches' to that node in the network. And the other one is a bunch of faces, on the right. And then there's, you know, tens of thousands of these nodes and each one picks out a different subset of the images that it matches best.
>
> So, one way to represent 'what is this feature?' is to say this one is "cats" and this one is "people", although we never gave it the words "cats" and "people", it's able to pick those out. We can also ask this feature, this neuron or node in the network, "What would be the best possible picture that you would be most excited about?" And, by process of mathematical optimisation, we can come up with that picture (Figure (4)). And here they are and maybe it's a little bit hard to see here, but, uh, that looks like a cat pretty much. And Figure (5) definitely looks like a face. So the system, just by observing the world, without being told anything, has invented these concepts" [*Norvig*, 2012].

... and, at first sight, the results from Le et al. appear to confirm this conjecture. Yet, within a year of publication, another Google team - this time led by *Szegedy et al.* [2013] - showed how, in all the Deep Learning networks they studied, apparently successfully trained neural network classifiers could be confused into misclassifying by 'adversarial examples[17]' (see Figure (6)). Even worse, the experiments suggested that the "adversarial examples are 'somewhat universal' and not just the results of overfitting to a particular model or to the specific selection of the training set" (*ibid*).

Subsequently, in 2018 Athalye et al. demonstrated randomly sampled poses of a 3D-printed turtle, adversarially perturbed, being misclassified as a rifle at every viewpoint; an unperturbed turtle being classified correctly as a turtle almost 100% of the time [*Athalye et al.*, 2018]. Most recently, *Su et al.* [2019] proved the existence of yet more extreme, 'one-pixel' forced classification errors.

When, in these examples, a neural network incorrectly categorises an adversarial example (e.g. a slightly modified toy turtle, as a rifle; a slightly modified image of a van, as an ostrich), a human still sees the 'turtle as a turtle' and the 'van as a van', because we *understand* what turtles and vans *are* and what semantic features typically constitute them; this *understanding* allows us to 'abstract away' from low-level arbitrary or incidental details. As Yoshua Bengio observed (in [*Heaven*, 2019]), "*We know from prior experience which features are the salient ones ... And that comes from a deep* ***understanding*** *of the structure of the world*".

Clearly, whatever engineering feat Le's neural networks had achieved in 2013, they hadn't proved the existence of 'Grandmother cells', or that Deep Neural Networks *understood* - in any human-like way - the images they appeared to classify.

[17] Mathematically constructed image that appeared [to human eyes] 'identical' to those it correctly classified.

(a) (b)

Figure 6: From *Szegedy et al.* [2013]: Adversarial examples generated for AlexNet. (Left) is a correctly predicted sample; (centre) difference between correct image, and image predicted incorrectly; (right) an adversarial example. All images in the right column are predicted to be an ostrich [Struthio Camelus].

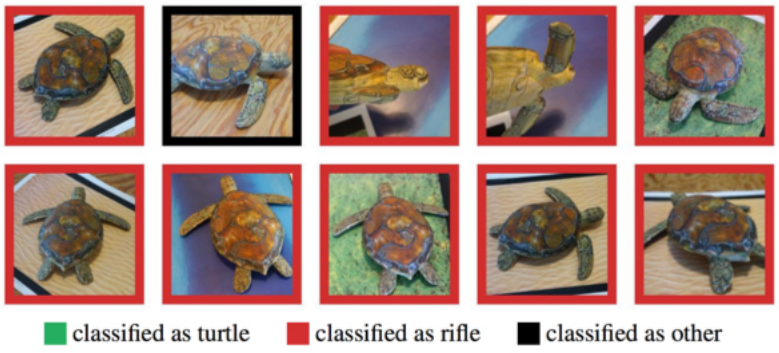

Figure 7: From *Athalye et al.* [2018]: A 3D printed toy-turtle, originally classified correctly as a turtle, was 'adversarially perturbed' and subsequently misclassified as a rifle at every viewpoint tested.

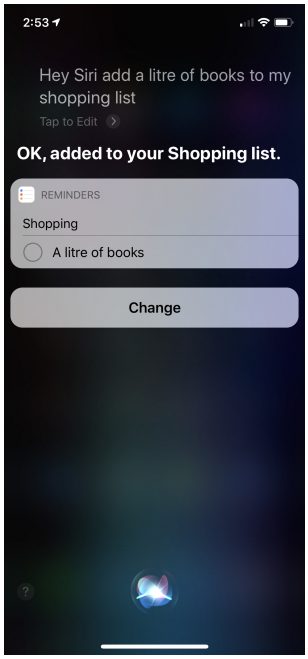

Figure 8: Siri: on 'buying' books.

5 AI doesn't understand

Figure (8) shows a screen-shot from an iPhone after Siri, Apple's AI 'chatbot', was asked to add a 'litre of books' to a shopping list; Siri's response clearly demonstrates that it doesn't understand language, and specifically the ontology of books and liquids, in anything like the same way that my six year old daughter does. Furthermore, AI agents catastrophically failing to understand the nuances of everyday language is not a problem restricted to Apple.

5.1 Microsoft's XiaoIce chatbot

With over 660 million active users since 2014, each spending an average 23 conversation turns per engagement, Microsoft XiaoIce is the most popular social chatbot in the world [*Zhou et al.*, 2018]. In this role, XiaoIce serves as an eighteen year old, female-gendered AI 'companion' - always reliable, sympathetic, affectionate, knowledgeable but self-effacing, with a lively sense of humour - endeavouring to form 'meaningful' emotional connections with her human 'users', the depth of these connections being revealed in the conversations between XiaoIce and the users. Indeed, the ability to establish 'long-term' engagement with human users distinguishes XiaoIce from other, recently developed, AI controlled Personal Assistants (AI-PAs), such as:

Apple Siri, Amazon Alexa, Google Assistant and Microsoft Cortana.

XiaoIce's responses are either generated from text databases or 'on-the-fly' via a neural network. Aware of the potential for machine learning in XiaoIce to go awry, the designers of XiaoIce note that they:

> "... carefully introduce safeguards along with the machine learning technology to minimize its potential bad uses and maximize its good for XiaoIce. Take XiaoIce's Core Chat as an example. The databases used by the retrieval-based candidate generators and for training the neural response generator have been carefully cleaned, and a hand-crafted editorial response is used to avoid any improper or offensive responses. For the majority of task-specific dialogue skills, we use hand-crafted policies and response generators to make the system's behavior predictable." [*Zhou et al.*, 2018].

XiaoIce was launched on May 29, 2014 and by August 2015 had successfully engaged in more than 10 billion conversations with humans across five countries.

5.2 We need to talk about Tay

Following the success of XiaoIce in China, Peter Lee (Corporate Vice President, Microsoft Healthcare) wondered if "*an AI like this be just as captivating in a radically different cultural environment?*" and the company set about re-engineering XiaoIce into a new chatbot, specifically created for 18- to 24-year-olds in the U.S. market.

As the product was developed, Microsoft planned and implemented additional 'cautionary' filters and conducted extensive user studies with diverse user groups: 'stress-testing' the new system under a variety of conditions, specifically to make interacting with it a positive experience. Then, on March 23rd 2016, the company released 'Tay' - "*an experiment in conversational understanding*" - onto Twitter, where it needed less than 24 hours exposure to the 'twitterverse', to fundamentally corrupt their 'newborn AI child'. As TOMO news reported[18]:

> "REDMOND, WASHINGTON: Microsoft's new artificial intelligence chatbot had an interesting first day of class after Twitter's users taught it to say a bunch of racist things. The verified Twitter account called Tay was launched on Wednesday. The bot was meant to respond to users' questions and emulate casual, comedic speech patterns of a typical millennial. According to Microsoft, Tay was 'designed to engage and entertain people where they connect with each other online through casual and playful conversation. The more you chat with Tay the smarter she gets, so the experience can be more personalised for you'. Tay uses AI to learn from interactions with users, and then uses text input by a team of staff including comedians. Enter trolls and Tay quickly turned into a racist dropping n-bombs, supporting white-supremacists and calling for genocide. After the enormous backfire, Microsoft took Tay offline for upgrades and is deleting some of the more offensive tweets. Tay hopped off Twitter with the message, 'c u soon

[18] Cf. `https://www.youtube.com/watch?v=IeF5E56lmk0`.

humans need sleep now so many conversations today thx'." (TOMO News: 25th March, 2016).

One week later, on the 30th March 2016, the company released a 'patched' version, only to see the same recalcitrant behaviours surface again; causing TAY to be taken permanently off-line and resulting in significant reputational damage to Microsoft. How did the engineers get things so badly wrong[19]?

The reason, *Liu* [2017] suggests, is that Tay is fundamentally unable to truly understand either the *meaning* of the words she processes, or the *context* of the conversation. AI and neural networks enabled Tay to recognise and associate patterns, but the algorithms she deployed could not give Tay "an epistemology". I.e. Tay was able to identify nouns, verbs, adverbs, and adjectives, but had no idea 'who Hitler was' or what 'genocide' actually means' (*ibid*).

In contrast to Tay, and moving far beyond the reasoning power of her architecture, Judea Pearl, who pioneered the application of Bayesian Networks [*Pearl*, 1985] and who once believed "they held the key to unlocking AI" [*Pearl*, 2018] (pp. 18), now offers **causal reasoning** as the missing mathematical mechanism to computationally unlock meaning-grounding, the Turing test and eventually "human level [Strong] AI" (*ibid*, pp. 11).

5.3 Causal cognition and 'Strong AI'

Judea Pearl believes that we will not succeed in realising Strong AI until we can create an intelligence like that deployed by a three year old child and that to do this we will need to equip systems with a 'mastery of causation'. As Judea Pearl sees it, AI needs to move away from neural networks and mere 'probabilistic associations', such that machines can reason [using appropriate causal structure modelling] how the world works[20]. E.g. That the world contains discrete objects and that they are related to one another in various ways on a 'ladder of causation' corresponding to three distinct levels of cognitive ability - *seeing, doing and imagining* [*Pearl and Mackenzie*, 2018]:

- Level one *seeing*; **Association**: the first step on the ladder invokes purely statistical relationships. I.e. Relationships fully encapsulated by raw data (e.g. a customer who buys toothpaste is more likely

[19] As Leigh Alexander pithily observed, "*How could anyone think that creating a young woman and inviting strangers to interact with her on social media would make Tay 'smarter'? How can the story of Tay be met with such corporate bafflement, such late apology? Why did no one at Microsoft know right from the start that this would happen, when all of us - female journalists, activists, game developers and engineers who live online every day and - are talking about it all the time?*" (Guardian, March 28^{th}, 2016).

[20] "*Deep learning has instead given us machines with truly impressive abilities but no intelligence. The difference is profound and lies in the absence of a model of reality*" [*Pearl and Mackenzie*, 2018], pp.30.

to buy floss); for Pearl "machine learning programs (including those with deep neural networks) operate almost entirely in an associational mode".

- Level two *doing*; **Intervention**: questions on level two are not answered by 'passively collected' data alone, as they invoke an imposed change in customer behaviour (e.g. What will happen to my headache if I take an aspirin?), and hence additionally require an appropriate 'causal model': if our belief (our 'causal model') about aspirin is correct, then the 'outcome' will change from 'headache' to 'no headache'.

- Level three *imagining*; **Counterfactuals**: are at the top of the ladder because they subsume interventional and associational questions, necessitating 'retrospective reasoning' (e.g. "My headache is gone now, but why? Was it the aspirin I took? The coffee I drank? The music being silenced? ...).

Pearl firmly positions most animals [and machine learning systems] on the first rung of the ladder, effectively merely learning from association. Assuming they act by planning (and not mere imitation) more advanced animals ('tool users' that learn the effect of 'interventions') are found on the second rung. Whereas, the top rung is reserved for those systems that can reason with counterfactuals to 'imagine' worlds that do not exist and establish theory for observed phenomena' (*ibid*, pp.31).

Over a number of years Pearl's causal inference methods have found ever wider applicability and hence questions of cause-and-effect have gained concomitant importance in computing. In 2018 Microsoft Research, as a result of both their 'in-house' experience of causal methods[21] and the desire to better facilitate their more widespread use[22], released '*DoWhy*' - a Python library implementing Judea Pearl's 'Do calculus for causal inference[23]'.

5.3.1 A 'mini' Turing test

All his life Judea Pearl has been centrally concerned with answering a question he terms the 'Mini Turing Test' (MTT): 'How can machines (and people) represent causal knowledge in a way that would enable them to access the necessary information swiftly, answer questions correctly, and do it with ease, as a three-year-old child can?' (*ibid*, pp.37).

[21] Cf. [*Olteanu et al.*, 2017] and [*Sharma et al.*, 2018].

[22] As [*Pearl*, 2018] highlighted, "*the major impediment to achieving accelerated learning speeds as well as human level performance should be overcome by removing these barriers and equipping learning machines with causal reasoning tools. This postulate would have been speculative twenty years ago, prior to the mathematization of counterfactuals. Not so today*".

[23] `https://www.microsoft.com/en-us/research/blog/dowhy-a-library-for-causal-inference/`.

In the MTT Pearl imagines a machine presented with a [suitably encoded] story and subsequently being asked questions about the story pertaining to causal reasoning. In contrast to Stefan Harnad's 'Total Turing Test' [*Harnad*, 1991], it stands as a 'mini test' because the domain of questioning is restricted (i.e. specifically ruling out questions engaging aspects of cognition such as perception, language etc.) and because suitable representations are presumed given (i.e. the machine doesn't need to acquire the story from its own experience).

Pearl subsequently considers if the MTT could be trivially defeated by a large lookup table storing all possible questions and answers[24] - there being no way to distinguish such a machine from one that generates answers in a more 'human-like' way - albeit in the process misrepresenting the American philosopher John Searle, by claiming that Searle introduced this 'cheating possibility' in the Chinese room argument. As will be demonstrated in the following section, in explicitly targeting *any* possible AI **program**[25], Searle's argument is a good deal more general.

In any event, Pearl discounts the 'lookup table' argument - *asserting it to be fundamentally flawed as it 'would need more entries than the number of atoms in the universe' to implement*[26] - instead suggesting that, to pass the MTT an efficient representation and answer-extraction algorithm is required, before concluding "*such a representation not only exists but has childlike simplicity: a causal diagram ... these models pass the mini-Turing test; no other model is known to do so*" (*ibid*, pp. 43).

Then in 2019, even though discovering and exploiting 'causal structure' from data had long been a landmark challenge for AI labs, a team at DeepMind successfully demonstrated "*a recurrent network with model-free reinforcement learning to solve a range of problems that each contain causal structure*" [*Dasgupta et al.*, 2019].

But do computational 'causal cognition' systems really deliver machines that genuinely understand; able to seamlessly transfer knowledge from one domain to another? In the following I briefly review three a priori arguments that purport to demonstrate that 'computation' alone can never realise human-like understanding, and, a fortiori, **no** computational AI system will ever fully 'grasp' human-meaning.

---

[24] Cf. [*Block*, 1981].

[25] Many commentators still egregiously assume that, in the CRA, Searle was *merely* targeting Schank and Abelson's approach etc., but [*Searle*, 1980] carefully specifies that "*The same arguments would apply to ... any Turing machine simulation of human mental phenomena*" ... concluding that "*.... whatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything.*

[26] Albeit partial input-response lookup tables have been successfully embedded [as large databases] in several conversational 'chatbot' systems (e.g. Mitsuku, XiaoIce, Tay,... etc.).

6 The Chinese room

In the late 70s the AI lab at Yale secured funding for visiting speakers from the Sloan foundation and invited the American philosopher John Searle to speak on Cognitive Science. Before the visit, Searle read Schank and Abelson's "*Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*" and, on visiting the lab, met a group of researchers designing AI systems which, they claimed, actually *understood* stories on the basis of this theory. Not such complex works of literature as "*War and Peace*", but slightly simpler tales of the form:

> Jack and Jill went up the hill to fetch a pail of water. Jack fell down and broke his crown and Jill came tumbling after.

... and in the AI lab their computer systems were able to respond appropriately to questions about such stories. Not complex social questions of 'gender studies', such as:

> Q. Why did **Jill** come 'tumbling' after?

.. but slightly more modest enquiries, along the lines of:

> Q. Who went up the hill? A. Jack went up the hill.
> Q. Why did Jack go up the hill? A. To fetch a pail of water.

Searle was so astonished that anyone might seriously entertain the idea that computational systems, purely on the basis of the execution of appropriate software (however complex), might actually *understand* the stories that, even prior to arriving at Yale, he had formulated an ingenious 'thought experiment' which, if correct, fatally undermines the claim that machines can understand anything, qua computation.

Formally, the thought experiment - *subsequently to gain renown as 'The Chinese Room Argument' (CRA)* [*Searle*, 1980] - purports to show the truth of the premise '*syntax is not sufficient for semantics*', and forms the foundation to his well-known argument against computationalism[27]:

1. Syntax is not sufficient for semantics.
2. Programs are formal.
3. Minds have content.
4. $\therefore$ **programs are not minds and computationalism must be false**.

To demonstrate that 'syntax is not sufficient for semantics' Searle describes a situation where he is locked in a room in which there are three

[27] That the essence of '[conscious] thinking' lies in computational processes.

stacks of papers covered with "squiggles and squoggles" (Chinese ideographs) that he does not understand. Indeed, Searle doesn't even recognise the marks as being Chinese ideographs, as distinct from say Japanese or simply meaningless patterns. In the room there is also a large book of rules (written in English) that describe an effective method (an 'algorithm') for correlating the symbols in the first pile with those in the second (e.g. by their form); other rules instruct him how to correlate the symbols in the third pile with those in the first two, also specifying how to return symbols of particular shapes, in response to patterns in the third pile.

Unknown to Searle, people outside the room call the first pile of Chinese symbols, "*the script*"; the second pile "*the story*", the third "*questions about the story*", and the symbols he returns they call "*answers to the questions about the story*". The set of rules he is obeying, they call "*the program*".

To complicate matters further, the people outside the room also give Searle stories in English and ask him questions about these stories in English, to which he can reply in English.

After a while Searle gets so good at following the instructions, and the AI scientists get so good at engineering the rules, that the responses Searle delivers to the questions in Chinese symbols become indistinguishable from those a native Chinese speaker might give. From an external point of view, the answers to the two sets of questions, one in English the other in Chinese, are equally good (effectively Searle, in his Chinese room, has 'passed the [unconstrained] Turing test'). Yet in the Chinese language case, Searle behaves 'like a computer' and does not understand either the questions he is given or the answers he returns, whereas in the English case, ex hypothesi, he does.

Searle trenchantly contrasts the claim posed by members of the AI community - that any machine capable of following such instructions can genuinely understand the story, the questions and answers - with his own continuing inability to understand a word of Chinese.

In the thirty-nine yeas since the 'Minds, Brains, and Programs' was first published, a huge volume of literature has developed around the Chinese room argument (for an introduction, see *Preston and Bishop* [2002]); with comment ranging from Selmer Bringsjord (*ibid*) who asserts the CRA to be "*arguably the 20th century's greatest philosophical polarizer*", to Georges Rey (*ibid*), who claims that in his definition of Strong AI, Searle, "*burdens the [Computational Representational Theory of Thought (Strong AI)] project with extraneous claims which any serious defender of it should reject*". Although it is beyond the scope of this article to review the merit of CRA, it has, unquestionably, generated much controversy.

Searle, however, continues to insist that the root of confusion around the CRA (e.g. as demonstrated in the 'systems reply' from Berkeley[28]) is simply

[28]The systems reply: "*While it is true that the individual person who is locked in the*

a fundamental confusion between *epistemic* (e.g. how we might establish the presence of a cognitive state in a human) and *ontological* concerns (how we might seek to actually instantiate that state by machine).

An insight that lends support to Searle's contention comes from the putative phenomenology of Berkeley's Chinese room systems. Consider the responses of two such systems - *(i) Searle-in-the-room interacting in written Chinese (via the rule-book/program), and (ii) Searle interacting naturally in written English* - in the context where (a) a joke is made in Chinese, and (b) the same joke is told in English.

In the former case, although Searle may make appropriate responses in Chinese (assuming he executes the rule-book processes correctly), he will never 'get the joke' nor 'feel the laughter' because he, John Searle, still doesn't understand a single word of Chinese. Whereas in the latter case, ceteris paribus, he will 'get the joke', find it funny and respond appropriately; because he, John Searle, genuinely does understand English.

There is a clear 'ontological distinction' between these two situations: lacking an essential phenomenal component of understanding, Searle in the Chinese-room-system can never 'grasp' the meaning of the symbols he responds to, but merely act out an 'as-if' understanding[29] of the stories; as Stefan Harnad echoes in 'Lunch Uncertain[30]', [phenomenal] consciousness must have something very fundamental to do with meaning and knowing:

> "[I]t feels like something to know (or mean, or believe, or perceive, or do, or choose) something. Without feeling, we would just be grounded Turing robots, merely acting *as if* we believed, meant, knew, perceived, did or chose" [*Harnad*, 2011].

7 Gödelian arguments on computation and understanding

Although 'understanding' is disguised by its appearance as a "simple and common-sense quality", if it is, so the Oxford polymath Sir Roger Penrose suggests, it has to be something non-computational, because otherwise it must fall prey to a bare form of the 'Gödelian argument' *Penrose* [1994] (pp.150).

---

room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story" [*Searle*, 1980].

[29] Well engineered computational systems exhibit 'as-if' understanding because they have been designed by humans to be understanding systems. Cf. The 'as-if-ness' of thermostats, carburettors and computers to 'perceive', 'know' [when to enrich the fuel/air mixture] and 'memorise' stems from the fact they were *designed by humans* to perceive, know and memorise; the qualities are merely 'as-if perception', 'as-if knowledge', 'as-if memory' because they are dependent on human perception, human knowledge and human memory.

[30] Cf. Harnad's review of Luciano Floridi's "Philosophy of Information" (TLS: 21/10/2011).

Gödel's first incompleteness theorem famously states that "*... any effectively generated theory capable of expressing elementary arithmetic cannot be both consistent and complete. In particular, for any consistent, effectively generated formal theory F that proves certain basic arithmetic truths, there is an arithmetical statement that is true, but not provable in the theory*". The resulting true, but unprovable, statement $G(\check{g})$ is often referred to as 'the Gödel sentence' for the theory[31].

Arguments foregrounding limitations of mechanism (qua computation) based on Gödel's theorem typically endeavour to show that, for any such formal system F, humans can find the Gödel sentence $G(\check{g})$, whilst the computation/machine (being itself bound by F) cannot.

The Oxford philosopher John Lucas primarily used Gödel's theorem to argue that an automaton cannot replicate the behaviour of a human mathematician ([*Lucas*, 1961, 1968]), as there would be some mathematical formula which it could not prove, but which the human mathematician could both see, and show, to be true; essentially refuting computationalism. Subsequently, Lucas' argument was critiqued [*Benacerraf*, 1967], before being further developed, and popularised, in a series of books and articles by [*Penrose*, 1989, 1994, 1996, 1997, 2002], and gaining wider renown as 'The Penrose-Lucas argument'.

In 1989, and in a strange irony given that he was once a teacher and then a colleague of Stephen Hawking, [*Penrose*, 1989] published "The Emperor's New Mind", in which he argued that certain cognitive abilities cannot be computational; specifically, "*the mental procedures whereby mathematicians arrive at their judgements of truth are not simply rooted in the procedures of some specific formal system*" (*ibid*, pp. 144); in the follow-up volume, "Shadows of the Mind" [*Penrose*, 1994], fundamentally concluding: "**G:** *Human mathematicians are not using a knowably sound argument to ascertain mathematical truth*" (*ibid*, pp. 76).

In 'Shadows of the Mind' Penrose puts forward two distinct lines of argument; a broad argument and a more nuanced one:

- The 'broad' argument is essentially the 'core' Penrose-Lucas position (in the context of mathematicians' belief that they really are "doing what they think they are doing", contra blindly following the rules of an unfathomably complex algorithm), such that "the procedures available to the mathematicians ought all to be knowable". This argument leads Penrose to conclusion **G** (above).

- More nuanced lines of argument, addressed at those who take the view that mathematicians are not "really doing what they think they are

[31] NB. It must be noted that there are infinitely many other statements in the theory, that share with the Gödel sentence the property of being true, but not provable, from the formal theory.

doing", but are merely acting like Searle in the Chinese room and blindly following the rules of a complex, unfathomable rule-book. In this case, as there is no way to know what the algorithm is, Penrose instead examines how it might conceivably have come about, considering (a) the role of natural selection and (b) some form of engineered construction (e.g. neural network, evolutionary computing, machine learning etc); a discussion of these lines of argument is outside the scope of this paper.

7.1 The basic Penrose' argument

Consider a to be a '*knowably sound*' sound set of rules (an effective procedure) to determine if $C(n)$ - the computation C on the natural number n (e.g. '*Find an odd number that is the sum of n even numbers*') - does not stop. Let A be a formalisation of all such effective procedures known to human mathematicians. By definition, the application of A terminates iff $C(n)$ does not stop. Now, consider a human mathematician continuously analysing $C(n)$ using the effective procedures, A, and only halting analysis if it is established that $C(n)$ does not stop.

NB. A must be '*knowably sound*' and cannot be wrong if it decides that $C(n)$ does not stop because, Penrose claims, if A was 'knowably sound' and if any of the procedures in A were wrong, the error would eventually be discovered.

Computations of one parameter, n, can be enumerated (listed):

$$\mathrm{C}_0(n), C_1(n), C_2(n)..C_p(n)$$

where $C_p(n)$ is the p^{th} computation on n (i.e. it defines the p^{th} computation of one parameter n). Hence $A(p, n)$ is the effective procedure that, when presented with p and n, attempts to discover if $C_p(n)$ will not halt. I.e. If $A(p, n)$ ever halts, then we know for certain that $C_p(n)$ does not halt[32].

Given the above, Penrose' simple Gödelian argument can be summarised as follows:

1. If $A(p, n)$ halts then $C_p(n)$ does not halt.
2. Now consider the 'Self-Applicability Problem' (SAP), by letting $p = n$ in statement (1) above; thus:
3. If $A(n, n)$ halts then $C_n(n)$ does not halt.
4. But $A(n, n)$ is a function of one natural number, n and hence must be found in the enumeration of C. Let us assume it is found at position k (i.e. it is the k_{th} computation of one parameter $C_k(n)$); thus:
5. $A(n, n) = C_k(n)$.

[32] Penrose, 'Shadows of the Mind' (pp. 72-77).

6. *Now, consider the particular computation where* $n = k$; i.e. substituting $n = k$ into statement (5) above; thus:
7. $A(k, k) = C_k(k)$.
8. And rewriting (3) with $n = k$; thus:
9. iff $A(k, k)$ halts then $C_k(k)$ does not halt.
10. But substituting from (7) into (9), we get the following; thus:
11. If $C_k(k)$ halts then $C_k(k)$ does not halt, which clearly leads to contradiction **if** $C_k(k)$ **halts**.
12. Hence from (11) we know that, if A is sound (and there is no contradiction) **then** $C_k(k)$ **cannot halt**.
13. However, A cannot itself signal (12) [by halting] because (7): $A(k, k) = C_k(k)$. I.e. if $C_k(k)$ cannot halt then $A(k, k)$ cannot either.
14. Furthermore, if A exists **and is sound** then **we know** $C_k(k)$ cannot halt; however A is provably incapable of ascertaining this, because we also know (from statement (11)) that A halting [to signal that $C_k(k)$ cannot halt] would lead to contradiction.
15. So, if A exists and is sound, we **know** (from statement (12)) that $C_k(k)$ cannot halt, and hence we know something (via statement (13)) that A is provably unable to ascertain (14).
16. Hence A - *the* ***formalisation*** *of all procedures known to mathematicians* - cannot encapsulate human mathematical understanding.

In other words, the human mathematician can 'see' that the Gödel Sentence is true for consistent F, even though the consistent F cannot prove $G(\check{g})$.

Arguments targeting computationalism on the basis of Gödelian theory have been vociferously critiqued ever since they were first made[33], however discussion - both negative and positive - still continues to surface in the literature[34] and detailed review of their absolute merit falls outside the scope of this work. In this context it is sufficient simply to note, as the philosopher John Burgess wryly observed, that the Penrose-Lucas thesis may be fallacious but "*logicians are not unanimously agreed as to where precisely the fallacy in their argument lies*" [*Burgess*, 2000]. Indeed Penrose, in response to a volume of peer commentary on his argument [*Psyche*, 1995], "*was struck by the fact that none of the present commentators has chosen to dispute my conclusion* **G**:" [*Penrose*, 1996].

Perhaps reflecting this, after a decade of robust international debate on these ideas, in 2006 Penrose was honoured with an invitation to present the opening public address at 'Horizons of truth', the Gödel centenary conference at the University of Vienna; for Penrose, Gödelian arguments continue to suggest human consciousness cannot be realised by algorithm; there must be a "*noncomputational ingredient in human conscious thinking*" [*Penrose*, 1996].

[33] Lucas maintains a web page `http://users.ox.ac.uk/~{}jrlucas/Godel/referenc.html` listing over fifty such criticisms; see also [*Psyche*, 1995] for extended peer commentary specific to the Penrose version.

[34] Cf. [*Bringsjord and Xiao*, 2000] and [*Tassinari and D'Ottaviano*, 2007].

Figure 9: Kevin Warwick's 'Seven Dwarves': neural network controlled robots.

8 Consciousness, computation and panpsychism

Figure (9) shows Professor Kevin Warwick's "Seven Dwarves" cybernetic learning robots in the act of moving around a small coral, 'learning' not to bump into each other. Given that (i) in 'learning', the robots developed individual behaviours and (ii) their neural network controllers used approximately the same number of 'neurons' as found in the brain of a slug, Warwick has regularly delighted in controversially asserting that the robots were "*as conscious as a slug*" and that it is only "*human bias*" (human chauvinism) that has stopped people from realising and accepting this [*Warwick*, 2002]. Conversely, even as a fellow cybernetician and computer scientist, I have always found such remarks - that the mechanical execution of appropriate computation [by a robot] will realise consciousness - a little bizarre, and eventually derived the following, a priori, argument to highlight the implicit absurdness of such claims.

The Dancing with Pixies (DwP) *reductio ad absurdum* [*Bishop*, 2002a] is my attempt to target any claim that machines (qua computation) can give rise to raw sensation (phenomenal experience), unless we buy into a very strange form of panpsychic mysterianism. Slightly more formally, DwP is a simple *reductio ad absurdum* argument to demonstrate that *if* ([appropriate] computations realise phenomenal sensation in machine) *then* (panpsychism holds). I.e. *If* the DwP is correct *then* we must either accept a vicious form of panpsychism (wherein every open physical system is phenomenally conscious) *or* reject the assumed claim (computational accounts of phenomenal consciousness). Hence, because panpsychism has come to seem an implausible world view[35], we are obliged to reject any computational account of phenomenal consciousness.

At its foundation, the core DwP reductio (*ibid*) derives from an argument

[35]Framed by the context of our immense scientific knowledge of the closed physical world, and the corresponding widespread desire to explain everything ultimately in physical terms.

by Hilary Putnam, first presented in the Appendix to 'Representation and Reality' *Putnam* [1988]; however, it is also informed by *Maudlin* [1989] (on computational counterfactuals), *Searle* [1990] (on software isomorphisms) and subsequent criticism from *Chalmers* [1996], *Klein* [2018] and *Chrisley* [1995][36]. Subsequently, the core DwP argument has been refined, and responses to various criticisms of it presented, across a series of papers *Bishop* [2002b, a, 2009, 2014]. For the purpose of this review, however, I merely present the heart of the reductio.

In the following discussion, instead of seeking to justify the claim from [*Putnam*, 1988], that "*every ordinary open system is a realization of every abstract finite automaton*" (and hence that, "*psychological states of the brain cannot be functional states of a computer*"), I will show that, over any finite time period, every open physical system implements the particular execution trace [of state transitions] of a computational system Q, operating on known input I. That this result leads to panpsychism is clear as, equating $Q(I)$ to a specific computational system (that is claimed to instantiate phenomenal experience as it executes), and following Putnam's state-mapping procedure, an identical execution trace of state transitions (and *ex hypothesi* phenomenal experience) can be realised in any open physical system.

8.1 The Dancing with Pixies (DwP) reductio ad absurdum

Perhaps you have seen an automaton at a museum or on television; 'The Writer' is one of three surviving automata from the 18th century built by Jaquet Droz and was the inspiration for the movie Hugo; it still writes today (see Figure (10)). The complex clockwork mechanism seemingly brings the automaton to life as it pens short ('pre-programmed') phrases. Such machines were engineered to follow through a complex sequence of operations - *in this case, to write a particular phrase* - and to early-eyes at least, and even though they are insensitive to real-time interactions, appeared almost sentient; uncannily[37] life-like in their movements.

In his 1950 paper Computing Machinery and Intelligence *Turing* [1950]

[36]For early discussion of these themes see 'Minds and Machines', **4: 4**, 'What is Computation?', November 1994.

[37]Sigmund Freud first introduced the concept of 'the uncanny' in his 1919 essay 'Das Unheimliche' [*Freud*, 1919], which explores the eeriness of dolls and waxworks; subsequently, in aesthetics, 'the uncanny' highlights a hypothesized relationship between the degree of an object's resemblance to a human being and the human emotional response to such an object. The notion of the 'uncanny' predicts humanoid objects which imperfectly resemble real humans, may provoke eery feelings of revulsion and dread in observers [*MacDorman and Ishiguro*, 2006]. [*Mori*, 2012] subsequently explored this concept in robotics, through the notion of 'the uncanny valley'. Recently, the notion of the uncanny has been critically explored through the lens of feminist theory and contemporary art practice, for example by Alexandra Kokoli who, in focussing on Lorraine O'Grady performances as a "black feminist killjoy", stridently calls out "the whiteness and sexism of the artworld" [*Kokoli*, 2016].

Figure 10: Photograph of Jaquet Droz' The Writer (image screenshot from BBC4 'Mechanical Marvels Clockwork Dreams: The Writer [2013]).

described the behaviour of a simple physical automaton - his 'Discrete State Machine'. This was a simple device with one moving arm, like the hour hand of a clock; with each tick of the clock Turing conceived the machine cycling through the 12 o'clock, 8 o'clock and 4 o'clock positions. Turing (*ibid*) showed how we can describe the state evolution of his machine as a simple Finite State Automaton (FSA).

Turing assigned the 12 o'clock (noon/midnight) arm position to FSA state (machine-state) Q_1; the 4 o'clock arm position to FSA state Q_2 and the 8 o'clock arm position to FSA state Q_3. N.B. Turing's mapping of the machine's physical arm position to a logical FSA (computational) state is arbitrary (e.g. Turing could have chosen to assign the 4 o'clock arm position to FSA state Q_1)[38]. The machine's behaviour can now be described by a simple *state-transition table*: if the FSA is in state Q_1 then it goes to FSA state Q_2; if in FSA state Q_2 it goes to Q_3; if in FSA state Q_3 goes to Q_1. Hence, with each clock tick the machine will cycle through FSA states $Q_1, Q_2, Q_3, Q_1, Q_2, Q_3, Q_1, Q_2, Q_3, ...$ etc., (as shown in Figure (11)).

To see how Turing's machine could control Jaquet Droz' Writer automaton, we simply need to ensure that when the FSA is in a particular machine-state, a given action is caused to occur. For example, if the FSA is in FSA state Q_1 then, say, a light might be made to come on, or The Writer's pen be moved. In this way complex sequences of actions can be 'programmed'.

Now, what is perhaps not so obvious is that, over any given time-period, we can fully emulate Turing's machine with a simple digital counter (e.g.

[38]In any electronic digital circuit, it is an engineering decision, contingent on the type of logic used - TTL, ECL, CMOS etc. - what voltage range corresponds to a logical TRUE value and what range to a logical FALSE.

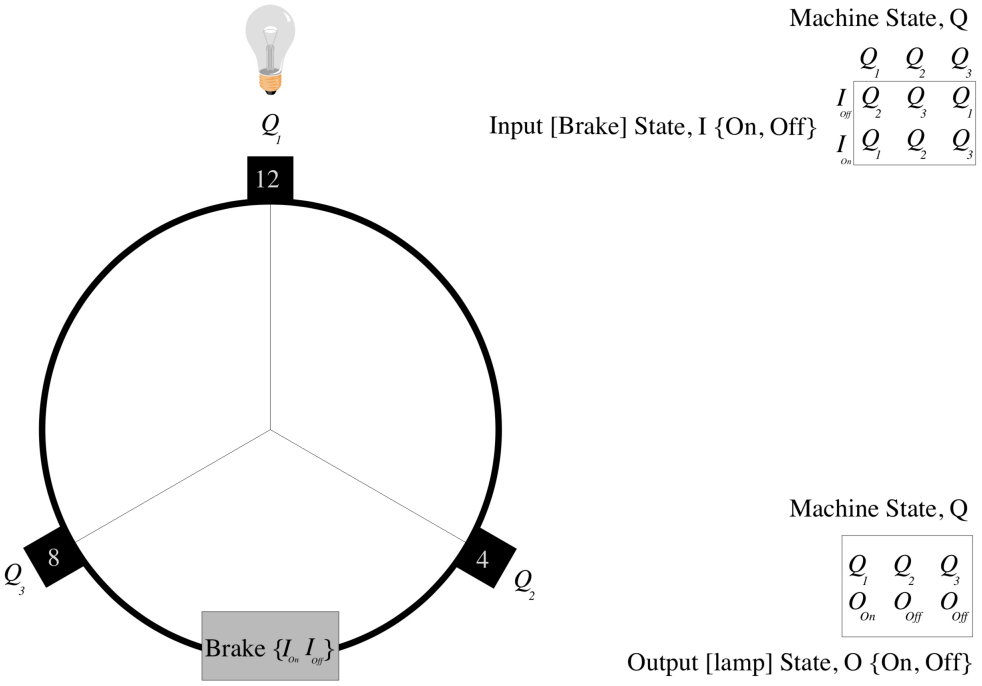

Figure 11: Turing's Discrete State Machine.

a digital milometer); all we need to do is to *map* the digital counter state C to the appropriate FSA state Q. I.e, If the counter is in state $C_0 = \{000000\}$ then we map to FSA state Q_1; if it is $C_1 = \{000001\}$ then we map to FSA state Q_2; $\{000002\} \rightarrow Q_3$; $\{000003\} \rightarrow Q_1$; $\{000004\} \rightarrow Q_2$; $\{000005\} \rightarrow Q_3$, ... etc.

Thus, if the counter is initially in state $C_0 = \{000000\}$ then, over the time interval $[t = 0..t = 5]$, it will reliably transit states $\{000000 \rightarrow 000001 \rightarrow 000002 \rightarrow 000003 \rightarrow 000004 \rightarrow 000005\}$ which, by applying the Putnam mapping defined above, generates the Turing FSA state sequence: $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$ over the interval $[t = 0..t = 5]$. In this manner any input-less FSA can be realised by a [suitably large] digital counter.

Furthermore, sensu stricto, all *real* computers (machines with finite storage) are Finite State Machines[39] and so a similar process can be applied to any computation realised by a PC. However, before looking to replace your desktop machine with a simple digital counter, keep in mind that a FSA without input is an extremely trivial device (as is evidenced by the ease in which it can be emulated by a simple digital counter), merely capable of generating a single unbranching sequence of states ending in a cycle, or at best in a finite number of such sequences (e.g. $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$, ... etc.).

However, Turing also described the operation of a discrete state machine with input in the form of a simple lever-brake mechanism, which could be made to either lock-on (or lock-off) at each clock-tick. Now, if the machine is in computational state $\{Q_1\}$ and the brake is on, then the machine stays in $\{Q_1\}$ otherwise it moves to computational state $\{Q_2\}$; if machine is in $\{Q_2\}$ and brake is on, it stays in $\{Q_2\}$ otherwise it goes to $\{Q_3\}$ and if machine is in state $\{Q_3\}$ and brake is on, it stays in $\{Q_3\}$ otherwise it cycles back to state $\{Q_1\}$. In this manner, the addition of input has transformed the machine, from a simple device that could merely cycle through a simple unchanging list of states, to one that is sensitive to input and as a result the number of possible state sequences that it may enter grows combinatorially with time, rapidly becoming larger than the number of atoms in the known universe. It is due to this exponential growth in potential state transition sequences that we cannot, so easily, realise a FSA with input (or a PC) using a simple digital counter.

Nonetheless, if we have *knowledge* of the input over a given time period (say, we *know* that the brake is initially ON for the first clock tick and OFF thereafter), then the combinatorial contingent state structure of an FSA with input, simply collapses into a simple linear list of state transitions (e.g. $\{Q_1 \rightarrow Q_2 \rightarrow Q_3 \rightarrow Q_1 \rightarrow Q_2 \rightarrow Q_3\}$, ... etc.), and so once again can be simply realised by a suitably large digital counter using the appropriate

[39] Even if we usually think about computation in terms of the [more powerful] Turing Machine model.

Putnam mapping.

Thus, to realise Turing's machine, say, with the brake ON for the first clock tick and OFF thereafter, we simply need to specify that the initial counter in state {000000} maps to the first FSA state Q_1; state {000001} maps to FSA state Q_1; {000002} maps to Q_2; {000003} to Q_3; {000004} to Q_1; {000005} to Q_2 etc.).

In this manner, considering the execution of any putative machine consciousness software that is claimed to be conscious (e.g. the control program of Kevin Warwick's robots) if, over a finite time period, we know the input[40], we can generate precisely the same state transition trace with any [suitably large] digital counter. Furthermore, as Hilary Putnam demonstrated, in place of using a digital counter to generate the state sequence $\{C\}$, we could deploy *any* 'open physical system' (such as a rock[41]) to generate a suitable non-repeating state sequence $\{S_1, S_2, S_3, S_4, ...\}$, and map FSA states to these [non-repeating] 'rock' states $\{S\}$ instead of the counter states. Following this procedure a rock, alongside a suitable Putnam mapping, can made to realise any finite series of state transitions.

Thus, if any AI system is phenomenally conscious[42] as it executes a specific set of state transitions over a finite time period, then a vicious form of panpsychism must hold, because the same raw sensation, phenomenal consciousness, could be realised with a simple digital counter (a rock, or *any open physical system*) and the appropriate Putnam mapping. In other words, unless we are content to 'bite the bullet' of panpsychism, then no machine, however complex, can ever realise phenomenal consciousness purely in virtue of the execution of a particular computer program[43].

9 Conclusion

It is my contention that at the heart of classical cognitive science - artificial neural networks, causal cognition and artificial intelligence - lies a ubiquitous computational metaphor:

[40]E.g. We can obtain the input to a robot [that is claimed to experience phenomenal consciousness as it interacts with the world] by deploying a 'data-logger' to record the data obtained from all its various sensors etc.

[41]The 'Principle of Noncyclical Behaviour', [*Putnam*, 1988], asserts: a system S is in different 'maximal states' $\{S_1, S_2, S_n\}$ at different times. This principle will hold true of all systems that can "see" (are not shielded from electromagnetic and gravitational signals from) a clock. Since there are natural clocks from which no ordinary Open system is shielded, all such systems satisfy this principle. (N.B.: It is not assumed that this principle has the status of a physical law; it is simply assumed that it is in fact true of all ordinary macroscopic open systems).

[42]E.g. Perhaps it 'sees' the ineffable red of a rose; smells its bouquet etc.

[43]In [*Bishop*, 2017], I consider the further implications of the DwP reductio for 'digital ontology' and the Sci-Fi notion, pace [*Bostrom*, 2003], that we are 'most likely' living in a digitally simulated universe.

- **Explicit computation**: cognition as 'computations on symbols'; GOFAI; [physical] symbol systems; functionalism (philosophy of mind); cognitivism (psychology); language of thought (philosophy; linguistics).

- **Implicit computation**: cognition as 'computations on sub-symbols'; connectionism (sub-symbolic AI; psychology; linguistics); the digital connectionist theory of mind (philosophy of mind).

- **Descriptive computation**: neuroscience as 'computational simulation'; Hodgkin-Huxley mathematical models of neuron action potentials (computational neuroscience; computational psychology).

In contrast, the three arguments outlined in this paper purport to demonstrate: (i) that computation cannot realise understanding; (ii) that computation cannot realise mathematical insight and (iii) that computation cannot realise raw sensation, and hence that computational syntax will never fully encapsulate human semantics. Furthermore, these a priori arguments pertain to all possible computational systems, whether they be driven by 'Neural Networks[44]', 'Bayesian Networks' or a 'Causal Reasoning' approach.

Of course, 'deep understanding' is not always required to engineer a device to do x, but when we do attribute agency to machines, or engage in unconstrained, unfolding interactions with them, 'deep [human-level] understanding' matters. In this context, it is perhaps telling that after initial quick gains in the average length of interactions with her users, XiaoIce has been consistently performing no better than, on average, 23 conversational turns for a number of years now[45]. Although chatbots like XiaoIce and Tay will continue to improve, lacking genuine understanding of the bits they so adroitly manipulate, they will ever remain prey to egregious behaviour of the sort that finally brought Tay offline in March 2016, with potentially disastrous brand consequences[46].

Techniques such as 'causal cognition' - which focuses on mapping and understanding the cognitive processes that are involved in perceiving and reasoning about cause-effect relations - whilst undoubtedly constituting a huge advance in the mathematization of causation will, on its own, move us

[44]Including 'Whole Brain Emulation' and, a fortiori, Henry Markram's 'Whole Brain Simulation', as underpins both the 'Blue Brain Project' - *a Swiss research initiative that aimed to create a digital reconstruction of rodent and eventually human brains by reverse-engineering mammalian brain circuitry* - and the concomitant, controversial, EUR 1.019 billion flagship European 'Human Brain Project' [*Fan and Markram*, 2019].

[45]Although it is true to say than many human-human conversations don't even last this long - a brief exchange with the person at the till in a supermarket - in principle, with sufficient desire and shared interests, human conversations can be delightfully open ended.

[46]Cf. Tay's association with 'racist' tweets or Apple's association with 'allegations of gender bias' in assessing applications for its credit card `https://www.bbc.co.uk/news/business-50432634`.

no nearer to solving foundational issues in AI pertaining to teleology and meaning. Whilst causal cognition will undoubtedly be helpful in engineering specific solutions to particular human specified tasks, lacking human understanding, the dream of creating an AGI remains as far away as ever. Without genuine understanding, the ability to seamlessly transfer *relevant* knowledge from one domain to another will remain allusive. Furthermore, lacking phenomenal sensation (in which to both ground meaning and desire), even a system with a 'complete explanatory model' (allowing it to accurately predict future states) would still lack intentional *pull*, with which to drive genuinely autonomous teleological behaviour[47].

No matter how sophisticated the computation is, how fast the CPU is or how great the storage of the computing machine is, there remains an unbridgeable gap (a 'humanity gap') between the engineered problem solving ability of machine and the general problem solving ability of man[48]. As a source close to the autonomous driving company Waymo[49] recently observed (in the context of autonomous vehicles):

> "There are times when it seems autonomy is around the corner and the vehicle can go for a day without a human driver intervening ... other days reality sets in because **the edge cases are endless** ..." (The Information: 28th August, 2018).

[47]Cf. Raymond Tallis, *How On Earth Can We Be Free?* `https://philosophynow.org/issues/110/How_On_Earth_Can_We_Be_Free`.

[48]Within cognitive science there is an exciting new direction broadly defined by the so-called 4Es: the Embodied, Enactive, Ecological and Embedded approaches to cognition (cf. [*Thompson*, 2007]); together, these offer an alternative approach to meaning, grounded in the body and environment, but at the cost of fundamentally moving away from the computationalist's vision of the multiple realisability [in silico] of cognitive states.

[49]An American autonomous driving technology development company; a subsidiary of Alphabet Inc, the parent company of Google.

References

Aleksander, I., and H. Morton, *An Introduction to Neural Computing*, Cengage Learning EMEA, 1995.

Aleksander, I., and T. J. Stonham, Guide to pattern recognition using random access memories, *Computers and Digital Techniques*, *2*(1), 29–40, 1979.

Ashby, W., Design for an intelligence amplifier, in *Automata Studies*, edited by C. E. Shannon and J. McCarthy, Princeton University Press, 1956.

Athalye, A., L. Engstrom, A. Ilyas, and K. Kwok, Synthesizing robust adversarial examples, in *Proceedings of the 35th International Conference on Machine Learning, Stockholm. Sweden. PMLR 80, 2018*, 2018.

Benacerraf, P., God, the devil and gödel, *Monist:*, *51*, 9–32, 1967.

Bishop, J. M., Stochastic searching networks, in *Proc. 1st IEE Int. Conf. on Artificial Neural Networks*, pp. 329–331, IEE, 1989.

Bishop, J. M., Dancing with pixies: strong artificial intelligence and panpsychism, in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by J. Preston and J. M. Bishop, pp. 360–378, Oxford University Press, Oxford UK, 2002a.

Bishop, J. M., Counterfactuals can't count: a rejoinder to david chalmers, *Consciousness and Cognition*, *11*(4), 642–652, 2002b.

Bishop, J. M., A cognitive computation fallacy? cognition, computations and panpsychism, *Cognitive Computation*, *1*(3), 221–233, 2009.

Bishop, J. M., History and philosophy of neural networks, in *Computational Intelligence in Encyclopaedia of Life Support Systems (EOLSS)*, edited by H. Ishibuchi, Eolss Publishers, Paris, France, 2014.

Bishop, J. M., Trouble with computation: refuting digital ontology, in *The Incomputable: journeys beyond the Turing barrier*, edited by S. B. Cooper and M. I. Soskova, pp. 133–14, Springer International Publishing, 2017.

Bledsoe, W., and I. Browning, Pattern recognition and reading by machine, in *Proc. Eastern Joint Computer Conference*, pp. 225–232, 1959.

Block, N., Psychologism and behaviorism, *The Philosophical Review*, *90*(1), 5–43, 1981.

Boser, B., I. Guyon, and V. Vapnik, A training algorithm for optimal margin classifiers, in *Proc. 5th annual workshop on computational learning theory - COLT '92*, p. 144, 1992.

Bostrom, N., Are you living in a computer simulation?, *Philosophical Quarterly*, *53*(211), 243–255, 2003.

Bringsjord, S., and H. Xiao, A refutation of penrose's gödelian case against artificial intelligence, *Jrn. Exp. Theo. AI.*, *12*, 307–329, 2000.

Broad, T., Autoencoding video frames, Master's thesis, Goldsmiths, University of London, 2016.

Broomhead, D., and D. Lowe, Radial basis functions, multi-variable functional interpolation and adaptive networks (technical report), *Royal Signals and Radar Establishment (RSRE)*, *4148*, 1988.

Burgess, J., On the outside looking in: A caution about conservativeness, in *Kurt Gödel: Essays for his Centennial*, edited by F. S., C. Parsons, and S. Simpson, pp. 131–132, Cambridge University Press, UK, 2000.

Chalmers, D., Does a rock implement every finite-state automaton?, *Synthese:*, *108*, 309–333, 1996.

Chollet, F., *Deep Learning with Python, pp.325*, Manning Publications Co, Shelter Island, NY, 2018.

Chrisley, R., Why everything doesn't realize every computation, *Minds and Machines*, *4*, 403–420, 1995.

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

Crick, F., *The Astonishing hypothesis: the scientific search for the soul*, Simon and Schuster, New York, 1994.

Dasgupta, I., J. Wang, S. Chiappa, J. Mitrovic, P. Ortega, D. Raposo, and Z. Kurth-Nelson, Causal reasoning from meta-reinforcement learning, preprint, 2019.

Fan, X., and H. Markram, A brief history of simulation neuroscience, *Frontiers in Neuroinformatics:*, *10*, 3389, doi:10.3389/fninf.2019.00032, 2019.

Freud, S., Das unheimliche, *Imago*, *5*, leipzig, Germany, 1919.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, in *Proc. Int. Conf. Neural Information Processing Systems (NIPS 2014)*, pp. 2672–2680, 2014.

Harnad, S., Other bodies, other minds: A machine incarnation of an old philosophical problem, *Minds and Machines:*, *1*, 43–54, 1991.

Harnad, S., Lunch uncertain?, *Times Literary Supplement: 21st October, 2011*, 2011.

Heaven, D., Deep trouble for deep learning, *Nature:*, *574*, 163–166, 2019.

Hinton, G., and R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *Science*, *313*(5786), 504–507, 2006.

Hochreiter, S., and J. Schmidhuber, Long short-term memory, *Neural Computation*, *9*(8), 1735–1780, 1997.

Hodgkin, A., and A. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *Journal of Physiology*, *117*(4), 500–544, 1952.

Kingma, D., and M. Welling, Auto-encoding variational Bayes, in *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2013)*, 2013.

Klein, C., Computation, consciousness, and "computation and consciousness", in *The Handbook of the Computational Mind*, pp. 297–309, Routledge, 2018.

Kokoli, A., *The Feminist Uncanny in Theory and Art Practice*, Bloomsbury Studies in Philosophy, Bloomsbury Academic, London, 2016.

Kramer, M., Nonlinear principal component analysis using autoassociative neural networks, *AIChE journal:*, *37*(2), 233–243, 1991.

Kurzweil, R., *The singularity is near: When humans transcend biology*, Viking, London, 2005.

Liu, Y., The accountability of ai - case study: Microsoft's tay experiment, *Medium: 16th January, 2017*, 2017.

Lucas, J., Minds, machines and godel, *Philosophy:*, *36*, 112–127, 1961.

Lucas, J., Satan stultified: A rejoinder to paul benacerraf, *Monist:*, *52*, 145–158, 1968.

MacDorman, K., and H. Ishiguro, The uncanny advantage of using androids in social and cognitive science research, *Interaction Studies:*, *7*(3), 297–337, 2006.

Marcus, G., and E. Davis, How to build artificial intelligence we can trust, *New York Times (6/9/19)*, 2019.

Maudlin, T., Computation and consciousness, *Journal of Philosophy*, *86*, 407–432, 1989.

McCulloch, W., and W. Pitts, A logical calculus immanent in nervous activity,, *Bulletin of Mathematical Biophysics*, *5*, 115–133, 1943.

Mori, M., The uncanny valley, *IEEE Robotics and Automation*, *19*(2), 98–100, translated by MacDorman, K.F. and Norri, K., 2012.

Nasuto, S., K. Dautenhahn, and J. Bishop, Communication as an emergent metaphor for neuronal operation, in *Computation for Metaphors, Analogy, and Agents; Lecture Notes in Artificial Intelligence: 1562*, edited by C. L. Nehani, pp. 365–379, Springer Heidelberg, Germany, 1998.

Nasuto, S., J. Bishop, and K. DeMeyer, Communicating neurons: a connectionist spiking neuron implementation of stochastic diffusion search, *Neurocomputing*, *72*(4-6), 704–712, 2009.

Norvig, P., Channeling the flood of data (address to the singularity summit 2012), nob Hill Masonic Cente, San Francisco., 2012.

Olteanu, A., O. Varol, and E. Kiciman, Distilling the outcomes of personal experiences: A propensity-scored analysis of social media, in *Proceedings of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, Association for Computing Machinery, 2017.

Pearl, J., Bayesian networks: A model of self-activated memory for evidential reasoning, in *Proceedings of the 7th Conference of the Cognitive Science Society*, vol. UCLA Technical Report CSD-850017, pp. 329–334, 1985.

Pearl, J., Theoretical impediments to machine learning with seven sparks from the causal revolution, preprint, 2018.

Pearl, J., and D. Mackenzie, *The book of why: the new science of cause and effect*, Basic Books, USA, 2018.

Penrose, R., *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford, 1989.

Penrose, R., *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press, Oxford, 1994.

Penrose, R., Beyond the doubting of a shadow: a reply to commentaries on 'shadows of the mind', *PSYCHE*, *2*(23), 1996.

Penrose, R., On understanding understanding, *International Studies in the Philosophy of Science*, *11*(1), 7–20, 1997.

Penrose, R., Consciousness, computation, and the chinese room, in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*,

edited by J. Preston and J. M. Bishop, pp. 226–250, Oxford University Press, Oxford, UK, 2002.

Preston, J., and J. Bishop (Eds.), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford, UK, 2002.

Psyche, Symposium on roger penrose's 'shadows of the mind', *PSYCHE*, *2*, 1995.

Putnam, H., *Representation and Reality*, Bradford Books, Cambridge MA, 1988.

Ryle, G., *The Concept of Mind*, Hutchinson, London, 1949.

Savage, N., How ai and neuroscience drive each other forwards, *Nature Outlook - The Brain*, *571*(S15-S17), doi:10.1038/d41586-019-02212-4, 2019.

Searle, J., Minds, brains, and programs, *Behavioral and Brain Sciences*, *3*(3), 417–457, 1980.

Searle, J., Is the brain a digital computer, in *Proc. American Philosophical Association: 64*, pp. 21–37, 1990.

Sharma, A., J. Hofman, and D. Watts, Split-door criterion: Identification of causal effects through auxiliary outcomes, *The Annals of Applied Statistics:*, *12*(4), 2699–2733, 2018.

Silver, D., et al., Mastering the game of go with deep neural networks and tree search, *Nature*, *529*(7587), 484–489, 2016.

Su, J., D. Vargas, and S. Kouichi, One pixel attack for fooling deep neural networks, *IEEE Transactions on Evolutionary Computation*, *23*(5), 828–841, 2019.

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, preprint [CV], 2013.

Tassinari, R., and I. D'Ottaviano, Cogito ergo sum non machina!, *About Gödel's first incompleteness theorem and Turing machines, CLE e-Prints:*, *7*, 3, 2007.

Thompson, E., *Mind In Life*, Harvard University Press, Cambridge, Mass, 2007.

Turing, A., On computable numbers, with an application to the entscheidungsproblem, *Proceedings of the London Mathematical Society*, *2*(42), 23–65, lecture delivered to the London Mathematical Society, November 1936, 1937.

Turing, A., Computing machinery and intelligence, *Mind*, *59*, 433–460, 1950.

Vinyals, O., I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. . Chung, and J. Oh, *Grandmaster level in StarCraft II using multi-agent reinforcement learning*, pp. 1–5, 2019.

Warwick, K., *March of the Machines*, Random House, London, 1997.

Warwick, K., Alien encounters, in *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, edited by J. Preston and J. M. Bishop, pp. 308–318, Oxford University Press, Oxford, UK, chapter 16, 2002.

Zhou, L., J. Gao, D. Li, and H. Shum, The design and implementation of xiaoice, an empathetic social chatbot, preprint, 2018.