From Matter to Life

Information and Causality

Edited by
Sara Imari Walker
Paul C.W. Davies
George F.R. Ellis

Contents

PART FOUR COMPLEXITY AND CAUSALITY 197

PART FIVE FROM MATTER TO MIND 267

18

Machine learning and the questions it raises

G. Andrew D. Briggs and Dawid Potgieter

Machine learning is a lively academic discipline and a key player in the continuous pursuit for new technological developments. The editorial in the first issue of the journal Machine Learning, published in March 1986, described the discipline as that field of inquiry concerned with the processes by which intelligent systems improve their performance over time, while recognising that it was hard to be more specific than the central tendency of the field (Langley, 1986). A glossary of terms published in the same journal in 1998 refined this to: Machine Learning is the field of scientific study that concentrates on induction algorithms and on other algorithms that can be said to "learn" (Kohavi and Provost, 1998).

Thomas J. Watson, the brilliant salesman who from 1914 to 1956 oversaw the remarkable growth and success of IBM – serving as both CEO and Chairman, was famously quoted as saying in 1943, *"I think there is a world market for maybe five computers"*. With more than one billion computers now in use worldwide (Virki, June 23 2008), this quote is often referenced to illustrate how vastly their usefulness had been underestimated. No area of computer science is making progress more rapidly than machine learning, with computers being capable of tasks that were a few decades ago only mentioned in science-fiction stories. Watson brought to IBM from his previous employment his trademark motto "THINK". It would at the time have been reasonable for Watson to suppose that only humans could really THINK. While computers could surpass humans in adding, subtracting, multiplying and dividing, they were hardly thought of as being good at human tasks, such as playing chess, which required thinking. This begs the question, "What is thinking". In February 1996 World Chess Champion Garry Kasparov took on the IBM computer Deep Blue in Philadelphia. Even with the IBM engineers allowed to reprogram the computer between games, the World Champion won, but only just, losing one game, drawing two, and winning three. His victory was short lived. The following year he played a rematch. With the score even after the first five of six games, Kasparov allowed Deep Blue to commit a knight sacrifice, which wrecked his Caro-Kann defences and forced him to resign in fewer than twenty moves.

It might be thought that playing chess is the kind of human thinking that is well-adapted to a digital computer, in which a computer can employ techniques for which the human brain is ill suited. A task for which the human eye-brain combination is superbly suited is facial recognition. It is not hard to understand why such expertise should be useful as humans organised themselves into hunter-gather groups, or even earlier. It is also not hard to think why such a task should be extraordinarily difficult for a digital computer. There are countless variations of lighting, viewing angle, mouth expressions, and other parameters. The advances have been made possible by a combination of raw computing power, vast data sets, and clever algorithms. At Heathrow airport iris scans have been superseded by comparison of what the camera sees with the picture stored in an e-passport. This is faster and more reliable than an experienced immigration officer, and as secure as the fingerprint method still used by US immigration authorities. Even voice recognition, which for years was the source of endless frustration for victims of automated telephone systems, can now be more accurate and more versatile than a human listener. The rest of this chapter will reflect on two areas of timely enquiry. By drawing from selected findings in neuroscience, the first part will explore to what extent we might expect computer processors to mimic the information processing mechanisms of the brain, an endeavour often referred to as neuromorphic technology. Recent findings in this area may provide deeper insights into information transfer in the brain and how such processes relate to machine learning. The second part will reflect on deeper questions, practical, ethical, and philosophical, which will inevitably need to be addressed as the learning capacity of machines continue to surpass that of humans in more and more ways.

The human brain uses very similar learning mechanisms to the brains of other mammals. This chapter will include findings from rodents and non-human primates from which more experimental data have been collected.

The human brain is arguably the most complex object in the known universe, the only "pound of flesh" taking credit for trying to understand itself. The mechanisms by which brains process information have for decades inspired computer scientists. When describing brains in terms of computation, neurons are often likened to wires. This metaphor may be useful in its simplest form: just as electricity flows from one wire to the next, an electrical signal can propagate along one neuron and be passed to another. Such a metaphor can also, within limited context, be extended to describe neural circuits as electric circuits and the brain as a very clever computer. But, the metaphor eventually fails. Connectivity between two neurons is not always constant, as it is with wires, but can vary depending on several factors, including the frequency of neurotransmission. This feature, called synaptic plasticity (see Table 18.1 for definitions), allows neurons to store and process information at the same time, thus integrating both memory and processing power, which are handled separately in computers.

Engineers have optimistically tackled the problem of mimicking synaptic plasticity, and memristor-based devices have been able to achieve this to some extent (Li et al., 2014). The prospect of building synapse-like computational devices therefore seems hopeful, but is this enough to build a computer which works like a brain? And if not, then what other mechanisms must be mimicked to achieve such a goal? The following observations may illustrate why the answer to the former question is 'no', and offer some reflections on how we might explore the answer to the latter.

Neurotransmission involves more than 'on' or 'off' signals. The kind of synapse which a memristor-based device might mimic makes use of ionotropic receptors, which, when bound by a neurotransmitter, either excite or inhibit the electrical activity of a recipient neuron. However, biological neurons also make use of metabotropic receptors, which may have little or no immediate effect on electrical activity, but do affect the cell over longer timescales. When metabotropic receptors are bound by a neurotransmitter, the downstream effects are mediated through a wide range of mechanisms available to the cell, including signalling cascades, metabolic activity, and changes to gene regulation (Diagram 1). Neurotransmission through metabotropic receptors play a crucial role in information processing. Dopamine for example, a neurotransmitter implicated in learning, decision-making, and movement, acts at D1- and D2-type metabotropic receptors (Beaulieu et al., 2015).

Table 18.1 *Neuroscience Terms and Definitions*

Term	Definition
Dopamine	A compound present in the body as a neurotransmitter; dopamine neurotransmission is implicated in learning, decision-making, and movement
En passant varicosities	Swellings along the length of an axon which do not always form a synapse, but they can sometimes release a neurotransmitter
Globus Pallidus	Part of the forebrain, located beneath the cerebral cortex towards the middle of the brain
Ionotropic receptors	Receptors that act as ion channels across the neuronal cell membrane, thereby changing the neuron's electrical conductivity
Metabotropic receptors	Receptors that act through signal transduction mechanisms inside the cell; such mechanisms may involve changes to metabolic processes or gene expression
Neural activity	Electrical activity of a neuron or cluster or neurons
Neuron	A specialised cell that carries electrical impulses; neurons make up less than half of the cells in a human brain
Neurotransmission	The release of a chemical from a neuron in order to send a message to other cells
Striatum	Part of the forebrain, located in humans between the globus pallidus and the cerebral cortex
Synapse	A structured junction between two neurons consisting of a small gap across which neurotransmitters can diffuse
Synaptic Plasticity	A change in the likelihood or efficiency of neurotransmission due to the frequency of its prior occurrence or other factors
Somatodentritic	A subcellular region of a neuron which normally receives signals from external neurotransmitters without sending any signals; the soma contains the cell nucleus where DNA is stored and transcribed

Information spreads beyond the synapse. The traditional metaphor that likens neurons to wires relies on neurotransmission occurring at distinct isolated regions called synapses, acting on one recipient neuron without spreading significantly to others. However, neurotransmission can, in some cases, spill over the synapse to influence multiple neurons in the vicinity. Dopamine released in the striatum, for example, has

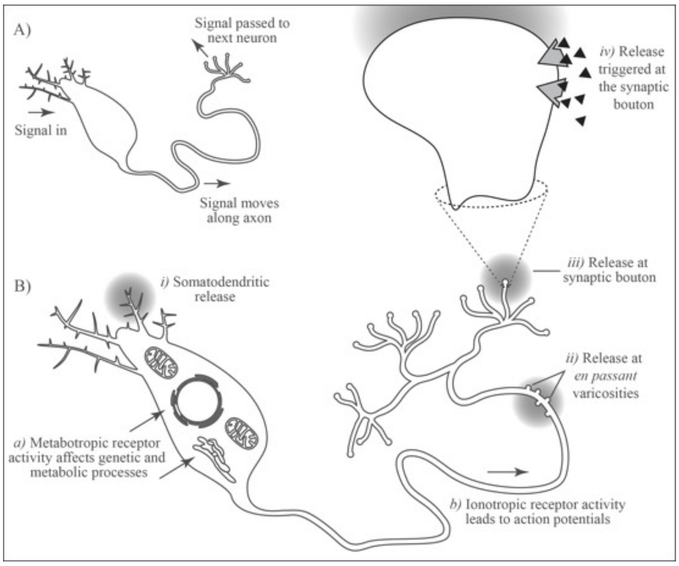
A)
Signal passed to next neuron
Signal in
Signal moves along axon
iv) Release triggered at the synaptic bouton
B)
i) Somatodendritic release
iii) Release at synaptic bouton
ii) Release at en passant varicosities
a) Metabotropic receptor activity affects genetic and metabolic processes
b) Ionotropic receptor activity leads to action potentials

Figure 18.1 A) Cartoon to illustrate the simplified model of neuron function. A signal is received at the dendrites or cell body, the signal propagates along an axon, and is passed to another neuron at the synaptic junction. B) An illustration of more complicated signalling mechanisms which are found in biological neurons. Activation of metabotropic receptors (a) leads to a variety of effects, including changes to gene regulation, protein expression, and metabolic processes. Activation of ionotropic receptors (b) leads to excitation or inhibition of action potentials. Neurotransmitter release has been observed from somatodendritic regions (i), and en passant along unmyelinated axons (ii). Different kinds of neurotransmitter release has been observed at synaptic boutons, including release which spill over significantly from the synaptic junction (iii), as well as release triggered at the synapse (iv) through activation of presynaptic membrane receptors.

been shown to diffuse into the extracellular space beyond the confines of the synapse (Cragg and Rice, 2004). Neurotransmission is also not limited to taking place at synaptic structures, but has been observed from somatodendritic parts of the neuron (Rice et al., 1997) and it has been suggested to take place at en passant varicosities along the length of an axon (Hattori et al., 1991) (See Fig. 1). Moreover, the axonal branches of a single dopamine neuron, the kind which might release dopamine from en passant varicosities, can spread out to innervate more than five per cent of the volume of striatum (Matsuda et al., 2009). This is because the axon branching is so extensive that if all branches of one such a neuron – from a rat brain – were connected end to end, the resulting structure would be nearly 80 cm in length (Matsuda et al. 2009). Taken together, these observations create a different perspective from one of the brain being like a collection of electrical wires connected end to end. Neurotransmitter release is more varied than that and can, in some neurons, take place beyond the confinement of synapses.

Information is coded at different levels – beyond the scale of the synapse. The traditional metaphor of the brain as a computer implies that information is embodied in an electrical signal which is passed from one neuron to another. However, information can affect different scales in the brain. For example, rhythmic neural activity, synchronised between multiple neurons in the globus pallidus, plays a crucial role in initiating goal-directed movement (Little and Brown, 2014). This kind of activity is composed of a collection of single action potentials, but it is not any action potential in particular, or even the sum of action potentials, but rather the synchronous and rhythmic nature of such action potentials, which plays a crucial role. Moreover, different frequencies of rhythmic activity, have been shown to play distinct roles in goal-directed actions (Brinkman et al., 2014). Such synchronous activity can therefore be thought of as an emergent phenomenon, and demonstrates that information processing within the brain can take place at different scales, which broadens the repertoire of information processing mechanisms available to the brain.

Neural circuits have large-scale redundancies. Computers and brains both employ error correction mechanisms, but the brain has a remarkable capacity to compensate functionally for the loss of neurons. Parkinson's disease is a degenerative condition in which the capacity to initiate goal-directed movements is affected. When a patient exhibits enough movement-related symptoms to be diagnosed, roughly half of the most vulnerable population of neurons have already died (Fearnley and Lees, 1991). By contrast, only a specially designed computer, such as the Hewlett-Packard Teramac built in 1997, would be immune to failure of even a small fraction of its components (Heath et al., 1998). The fault tolerance of Teramac however, relied on a 24-hour-long procedure during which a separate workstation identified all defective resources and wrote their locations to a configuration table as being 'in use' (Birnbaum and Williams, 2000). In essence, Teramac required the help of another computer to make sure it could start in spite of its faulty components. Brains are of course much more versatile in this sense. Unlike wires and transistors, which tend to need replacing when broken, the connections between neurons can change and grow dynamically without the need for invasive intervention. The biological mechanisms which underlie brain function can compensate for (a certain level of) failures whenever they occur. In the case of some forms of brain damage, such as the kind of cell death suffered during a stroke, the neurons surrounding a damaged area can 'rewire' themselves by sprouting new axon branches and making new connections to other cells (Dancause et al., 2005). In the case of degenerative diseases such as Parkinson's, the surviving neurons can increase the amount of dopamine output until they too become dysfunctional. Other neurons which normally release serotonin can also begin to release dopamine, albeit sometimes inappropriately (Tanaka et al., 1999).

Some neurons release more than one type of neurotransmitter. It has been suggested that some neurons release both excitatory and inhibitory neurotransmitters (Gutiérrez, 2005), which may increase the repertoire of possible signals being transmitted between neurons (Seal and Edwards, 2006). Although such a mechanism may not serve both to excite and to inhibit the same recipient cell, as doing so would be counter productive, it is feasible to imagine the neurotransmission from a single neuron exciting some recipient neurons and inhibiting others. Such a function could be achieved either by the recipient neurons being restricted in the receptors they express, or by the different neurotransmitters being located in different and adequately separated axonal branches. The capacity for neurons to release more than one neurotransmitter also extends the potential for different possible computations. Input to a single neuron does not lead to a single output or even many outputs of the same kind, but to multiple, and possibly a range of, different kinds of outputs.

Not all action potentials lead to neurotransmitter release. A single action potential contributes only to the probability of neurotransmitter release. For some neurons, there is a high probability that a single action potential will result in neurotransmitter release, but for others the probability can be very low, allowing additional factors to influence neurotransmission. For example, neurotransmission from dopamine axons in the striatum is influenced both by the frequency of action potentials and by the concentration of another neurotransmitter around the axon terminal, namely acetylcholine (Zhang and Sulzer, 2004) (See Diagram 1). The amount of dopamine released by these neurons is strongly proportional to the frequency of action potentials when acetylcholine is absent. The higher the rate of action potentials, the more dopamine will be released. However, when a significant concentration of acetylcholine is present, then the magnitude of dopamine release is changed such that action potentials at both low and high frequencies lead to a moderate amount of dopamine release. Moreover, the synchronised activation of acetylcholine neurons, when surrounding dopamine axon terminals, is sufficient to cause dopamine release without any action potentials being present at all in the dopamine neurons (Threlfell and et al, 2012). These findings suggest that some neurons behave almost like a logic gae, or rather – since the neurons in question also branch out across a relatively large volume of the striatum (Matsuda et al. 2009) – some neurons can behave like a collection of different logic gates. Whether such a collection acts in unison or whether each neuron acts independently is not completely understood. Taken together, the observations presented here testify that a simple model which describes the brain in terms of a bunch of interconnected electrical wires is out-dated, which may in turn suggest new directions for machine learning.

Our understanding of the brain is becoming increasingly sophisticated. Mammalian brains, and primate brains in particular, utilise an immense repertoire of mechanisms for processing information. Engineers who aim to build computers which utilise the same mechanisms as brains will have to continue developing more complex technologies, but taking on such a task for the purpose of building better computers may be both overambitious and unnecessary. Trying to understand every biological mechanism in the brain for the purpose of building a computer might be a bit like researching the type font and ink composition of a book simply to quote a paragraph from it. How then should scientists approach machine learning in relation to the human brain? The evidence presented here suggests three approaches. First, the different scales at which brains process information remain incompletely understood, and warrant continued investigation. How much of the information processing takes place through molecular interactions within a cell? How much takes place through electrical signals passed between individual neurons, and how much takes place through the synchronised rhythmic activity of large groups of neurons? Alongside research into scales of information processing, come questions about the hierarchy of algorithmic processes. Computers make use of a hierarchy of algorithmic processes, but the brain's capacity for algorithmic processes remains incompletely understood. Can the function of a healthy human brain can be reduced a set of district algorithmic processes, and how? Information is not only processed by electrochemical activity between neurons, but might also involve complex molecular interactions inside single neurons. It may therefore be difficult to determine where the boundaries lie between hierarchies of algorithms.

Second, we need a deeper understanding of the full repertoire of computational mechanisms available to the brain. More accurate models of brain function will provide a foundation constructing new models of neural processes. For example, a neuron which only releases glutamate, and only does so at synaptic structures, might be modelled as a wire with plasticity at the synapse – but a neuron which releases both glutamate (excitatory) and GABA (inhibitory) could be modelled as two circuits with the same inputs, but opposite output states and different output connections.

Finally, progress in machine learning may be made by looking at the function of biological mechanisms rather than aiming to build synthetic versions of the same hardware. For example, the memristor-based devices which were created as a synthetic alternative to synaptic plasticity (Li et al., 2014) will not necessarily be able to serve the function of a biological synapse. The latter is diverse, as biological mechanisms tend to be. Neurons can achieve plasticity through a range of signalling molecules binding at a variety of receptors, thereby triggering changes which include genetic expression, the relocation of proteins, and the regulation of their activity (Citri and Malenka, 2008). On the other hand, memristor-based devices – and semiconductor materials in general – tend to be much simpler in their construction and more limited in the variety of actions that are available at a molecular level. An individual neuron may not be just a component of a computation, but might perform a complete computation. Conversely, computations take place not just within individual neurons, but also between neurons, and sometimes, the activity of a single neuron is redundant. The constraints and resources available to biological organisms differ from those which are available in the laboratory. Electronic devices can perform similar information processing as brains by different – and possibly more efficient – mechanisms. The best goal may not be to build a computer with the same capacities as a brain, but rather to develop a range of machines, each capable of outperforming human brains in different ways.

Quantum Computers are utterly different from classical computers, and for certain tasks they can be exponentially better. At the core of machine learning lies probability theory, Bayesian theory in particular. To learn involves inference of probabilities over variables of interest and this is underpinned by two key operations; optimisation and numerical integration. Both scale poorly with the complexity of the models used and indeed with the amount of data. Classical computation cannot resolve this curse of dimensionality. Quantum computation offers an avoidance for such problems; indeed work in Information Engineering at the University of Oxford has shown that classical inference ground energy states in a quantum system are equivalent; furthermore quantum algorithms, inspired by classical sample-based inference, can also be developed to replace the most used classical inference algorithms (Fox et al., 2008). Unusually for quantum information processing, these algorithms depend for their effectiveness on decoherence to disperse unwanted information into surrounding coolants. If the algorithm provides a useful computational speedup then maybe evolution would have found a way to use it; it is an empirical question whether this is actually the case. There is a subneuronal computation theory that could be extended to perform this kind of MRF inference, making use of biological coolants flowing through microtubules to provide the necessary decoherence. Quantum methods have been proposed to reduce the vastness of the data sets which have proved crucial for machine learning, using superposition states not only of the information itself but also of the address switches used to retrieve data from classical memories (Lloyd et al., 2013). This would yield a logarithmic reduction in the amount of memory required to store the quantum information.

Having briefly reflected on which lines of enquiry might help us to build better machines, the remainder of this chapter will focus more broadly on a few related questions. Information is not just a set of bits, or even bits that convey semantic meaning, but information can have a causal power given the appropriate conditions. In the context of machine learning, the causal power of information is increasing dramatically, which can have significant implications for humanity. The industrial revolution saw a huge increase in our productivity through the power of machines. But humans were still very much in control. We now approach technological advances which allow for humans not necessarily having to be in control; machines are gaining autonomy. Although machines can process information well enough to become less reliant on human decisions, machines lack information for concepts like ‘good’ and ‘bad’. They do not have a framework for morals or values as we understand these concepts, or in fact, as philosophers still debate them. The questions discussed in this section explore some of these broader implications of recent advances in machine learning. The day may be not far off, indeed it may already be here, when machine learning advances to the point of creating new challenges to humans, for example, by undertaking tasks which traditionally we might have ascribed to human judgement. A machine might learn enough to be able to analyse the information in a company’s accounts to the standard of an experienced accountant. It might not only be able to answer the key question of an audit: are these accounts a true and accurate statement of the financial affairs of the organisation? It might be able to do more: what changes might increase the profitability or reduce the tax liability?

The rapid advances in information processing by machine learning raises questions which are better tackled sooner rather than later.

1 **What will be the implications for human work of machine learning?** Every previous advance in information processing, indeed in technology generally, has resulted in changes to human employment (Uglow, 2003). The mechanisation of agriculture led to urbanisation. The industrialisation of pottery making led to factory working. The results were a mixture of poverty and deprivation for the unemployed in cities, and steady wages and prosperity for many employed in factories. Nineteenth century England saw both at the same time. The computerisation of banking and financial services in the twentieth century removed the need for large numbers of people in clerical work. At its best it liberated them to use their brainpower for more interesting knowledge-based work. Will machine learning do that? Certain surgical operations can be performed better by a robot than even the best trained pair of human hands (Badani et al., 2007), thus enabling more accurate surgery to be performed and liberating the surgeon to concentrate on higher level tasks of diagnosis and treatment planning. Already people tend to be more honest about their potentially-embarrassing activities which led to sickness when answering questions for a computer rather than a clinician. What if the day comes when machines can outperform surgeons at diagnosis, treatment planning, and all surgical procedures? What higher level activity will then be left to the consultant? More positively, how can technological advancements such as these be leveraged to bring about the most good to society? Much of the work mentioned in this book, and elsewhere (Mirmomeni et al., 2014), explore the idea that information ? and in particular, having the correct information and the capacity process it well ? can be conceptualised as making a key contribution to the survival and evolution of an organism. Biological evolution of humans is a very slow process, and one with which we should not interfere. However, the capacity of machines to process information is growing rapidly, and vastly exceeds the rate of evolution of humans. Can society benefit from engaging with such technological advances as progress that benefits our species as a whole? Can such an approach inspire more people to create helpful technologies that will change the way that people work even more that the industrial revolution did in the 18th century?

2 **Who will set the hierarchy of goals for machine learning?** At present the programmer sets the goals: compare that face with the record in the e-passport; transcribe this speech into text. In 2014 TheySay, a spin out company from Oxford Uni-

versity, algorithmically analysed the sentiments of YES and NO supporters leading up to the referendum vote about whether Scotland should leave the United Kingdom (Morgan, March 10 2015). Harvesting text from social media, news, blogs and chat rooms, the machine used over 68,000 rules to determine the grammatical content of the text. It was then able to extract and assign meaning and provide insights about sentiment (positive or negative), emotions (fear, anger, happiness and sadness), and sureness (certainty and confidence), thus building up a picture of outrage, nervousness, despondency and joy. These goals were set by humans for the purposes which they wanted to achieve. What happens when powerful machines like these are programmed for maleficent purposes? The legal and ethical framework which is required to safely use new technology is sometimes developed much later than the technology itself. For example, one of the world's largest mass multiplayer online (MMO) computer games, Eve Online, has suffered losses of tens of thousands of US dollars because the technology was developed without a system to ensure ethical behaviour and justice. Some players stole online game credits from other players, which were then traded in for cash. The largest recorded theft was worth over US $51,000 (Drain, October 28 2012). How can we learn from this sort of dilemma? What might happen if we develop machines with the information processing power to set their own goals, but without the necessary framework to do so responsibly? Should goal-setting for machines be pre-emptively regulated? If so, how it should be done?

3 **What is the meaning of responsibility in machine learning?** The December 2014 issue of *Nature Nanotechnology* carried a thesis piece entitled *"Could we 3D print an artificial mind"*? (Maynard, 2014). The final paragraph concluded, *"Which leads us to a question that is, if anything, more difficult to address than the aforementioned technical hurdles: if our technological capabilities are beginning to shift from the fanciful to the plausible in constructing an artificial mind that has some degree of awareness, how do we begin to think about responsibility in the face of such audacity?"* One of us wrote to the author to ask him whether this final paragraph was simply a rhetorical flourish with which to end the piece, or whether these were issues about which he had been thinking deeply. He replied that his intent was to finish with a link to his broader work, which revolves around emerging technologies and responsible innovation. The responsibility he was referring to is that incumbent on the various societal actors who may be involved in the development and use of technologies that could lead to artificial minds or similar. He confessed that he had not thought about responsibility from the perspective of the artificial mind in the piece. He acknowledged that this is an intriguing avenue to go down, which touches on some of the current philosophical work around artificial general intelligence .

4 **Could machine learning constitute a threat to our existence?** The Oxford philosopher Professor Nick Bostrom thinks so. *"If machine brains surpassed human brains in general intelligence, then this new superintelligence could become extremely powerful – possibly beyond our control. As the fate of the gorillas now depends more on humans than on the species itself, so would the fate of humankind depend on the actions of the machine superintelligence?* (Bostrom, 2014). In January 2014 at an FQXi conference in Vieques on The Physics of Information, the director, Professor Max Tegmark, initiated a straw poll among the participants to ask what they considered to be at present the greatest risk to the survival of the human race [http://fqxi.org/conference/2014]. The obvious suspects, such as global climate change, came rather far down the list. Second from top was the threat from synthetic biology: the risk that some maverick, or perhaps some future child who had been given a gene splicing kit for Christmas, might produce a virus to which humans could

produce insufficient immunity. The highest threat was perceived to come from artificial general intelligence . Dilemmas in this area were foreseen as early as 1942 in Iszaac Azimov's short story, Runaround, in which machines have to obey three laws loosely paraphrased as: (1) never letting a human come to harm; (2) obeying orders, unless rule 2 conflicts with 1; (3) preserving itself, unless rule 3 conflicts with 1 or 2. In the story, a robot which was very expensive, and therefore programmed with a priority to preserve itself, is ordered to go on a dangerous mission which leads to a conflict between rules 2 and 3. The robot consequently malfunctions, and does so nearly to the peril of humans. That story was no more than fiction. Is the day approaching when such fiction might become reality? What if machines could not only be cleverer than humans but could adopt goals which were malevolent (again to beg the question) towards humans? So-called 'algorithmic self-improvement' might runaway and produce systems that greatly exceed human intelligence in ways that might not be human friendly. The long established philosophical tools of decision theory may be relevant for addressing questions of goal stability under algorithmic self-improvement. Could we run simulations on machine learning to predict whether such goals will evolve? If machine learning can in principle threaten human existence, then how can we find the best way to prevent such a threat? Could machine learning be programmed to act with character virtues such as humility, forgiveness, and kindness?

5 **Where is wisdom to be found for machine learning?** If machines can learn to be as intelligent as humans, can they also learn to be as wise? What would it mean to describe computing as wise? Would it involve being morally careful or perceptive? Would it make sense to ascribe attributes of wisdom to a machine only if one could also ascribe attributes of foolishness? At the meeting of the American Association for the Advancement of Science in San Jose in February 2015, a session was devoted to Wise Computing. Professor Kazuo Iwano, of the Japan Science and Technology Agency, introduced the session under its original title of Wisdom Computing (Iwano, 2015). He described how research activities in Japan are working to understand and develop wisdom by sublimating distributed and heterogeneous data and information. Humans are capable of accessing more information than ever in real time, but can we claim that we have become wiser than ever individually or collectively? Machine learning is attaining enormous capabilities in accessing and analyzing information and controlling objects such as airplanes and automobiles. Iwano presented a chart with the abscissa indicating on a logarithmic scale the duration of information, and the ordinate indicating the extent of its dissemination. Text messages scored poorly on both scales. Shakespeare scored highly on both. Top of all came the Bible. Could machines indeed acquire wisdom by learning from sources of spiritual information such as the ancient scriptures? If machines can learn to be wise, then what would that look like? Could 'machine wisdom' become an integral part of technology as machines gain greater capacity for performing new tasks, and outperforming humans?

References

Badani, K. K., Kaul, S., and Menon, M. 2007. Evolution of robotic radical prostatectomy. *Cancer*, **110**(9), 1951–1958.

Beaulieu, J. M., Espinoza, S., and Gainetdinov, R. R. 2015. Dopamine receptors–IUPHAR review 13. *British journal of pharmacology*, **172**(1), 1–23.

Birnbaum, J., and Williams, R. S. 2000. Physics and the information revolution. *Physics Today*, **53**(1), 38–43.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Brinkman, L., Stolk, A., Dijkerman, H. C., de Lange, F. P., and Toni, I. 2014. Distinct roles for alpha-and beta-band oscillations during mental simulation of goal-directed actions. *The Journal of Neuroscience*, **34**(44), 14783–14792.

Citri, A., and Malenka, R. C. 2008. Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology*, **33**(1), 18–41.

Cragg, S. J., and Rice, M. E. 2004. DAncing past the DAT at a DA synapse. *Trends in neurosciences*, **27**(5), 270–277.

Dancause, N., Barbay, S., Frost, S.B., Plautz, E.J., Chen, D., Zoubina, E.V., Stowe, A.M., and Nudo, R.J. 2005. Extensive cortical rewiring after brain injury. *The Journal of neuroscience*, **25**(44), 10167–10179.

Drain, B. October 28 2012. *EVE Evolved: Top Ten Ganks, Scams, Heists and Events*.

Fearnley, J. M., and Lees, A. J. 1991. Ageing and Parkinson's disease: substantia nigra regional selectivity. *Brain*, **114**(5), 2283–2301.

Fox, C., Rezek, L., and Roberts, S. 2008. *Local Quantum Computing for Fast Probably MAP Inference in Graphical Models*.

Gutiérrez, R. 2005. The dual glutamatergic–GABAergic phenotype of hippocampal granule cells. *Trends in neurosciences*, **28**(6), 297–303.

Hattori, T., Takada, M., Moriizumi, T., and Van der Kooy, D. 1991. Single dopaminergic nigrostriatal neurons form two chemically distinct synaptic types: possible transmitter segregation within neurons. *Journal of Comparative Neurology*, **309**(3), 391–401.

Heath, J. R., Kuekes, P. J., Snider, G. S., and Williams, R. S. 1998. A defect-tolerant computer architecture: Opportunities for nanotechnology. *Science*, **280**(5370), 1716–1721.

Iwano, K. 2015. *Wise Computing: Collaboration Between People and Machines*. In . San Jose, CA.

Kohavi, R., and Provost, F. 1998. Glossary of terms. *Machine Learning*, **30**(2-3), 271–274.

Langley, P. 1986. Editorial: On machine learning. *Machine Learning*, **1**(1), 5–10.

Li, Y., Zhong, Y., Zhang, J., Xu, L.and Wang, Q., Sun, H., Tong, H., Cheng, X., and Miao, X. 2014. Activity-dependent synaptic plasticity of a chalcogenide electronic synapse for neuromorphic systems. *Scientific reports*, **4**.

Little, S., and Brown, P. 2014. The functional role of beta oscillations in Parkinson's disease. *Parkinsonism & related disorders*, **20**, S44–S48.

Lloyd, S., Mohseni, M., and Rebentrost, P. 2013. Quantum algorithms for supervised and unsupervised machine learning. *arXiv preprint arXiv:1307.0411*.

Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., and Arai, R.and Kaneko, T. 2009. Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. *The Journal of Neuroscience*, **29**(2), 444–453.

Maynard, A. D. 2014. Could we 3D print an artificial mind? *Nature nanotechnology*, **9**(12), 955–956.

Mirmomeni, M., Punch, W. F., and Adami, C. 2014. Is information a selectable trait? *arXiv preprint arXiv:1408.3651*.

Morgan, D. March 10 2015. *The Zeitgeist in the Machine.*

Rice, M. E., Cragg, S. J., and Greenfield, S. A. 1997. Characteristics of electrically evoked somatodendritic dopamine release in substantia nigra and ventral tegmental area in vitro. *Journal of Neurophysiology*, **77**(2), 853–862.

Seal, R. P., and Edwards, R. H. 2006. Functional implications of neurotransmitter corelease: glutamate and GABA share the load. *Current opinion in pharmacology*, **6**(1), 114–119.

Tanaka, H., Kannari, K., Maeda, T., Tomiyama, M., Suda, T., and Matsunaga, M. 1999. Role of serotonergic neurons in L-DOPA-derived extracellular dopamine in the striatum of 6-OHDA-lesioned rats. *Neuroreport*, **10**(3), 631–634.

Threlfell, S., and et al. 2012. Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. *Neuron*, **75**(1), 58–64.

Uglow, J. 2003. *The lunar men: Five friends whose curiosity changed the world.* Macmillan.

Virki, T. June 23 2008. *Computers in Use Pass 1 Billion Mark.*

Zhang, H., and Sulzer, D. 2004. Frequency-dependent modulation of dopamine release by nicotine. *Nature neuroscience*, **7**(6), 581–582.