Chapter 1
On the simulation (and energy costs) of human intelligence, the singularity and simulationism

Alan F.T. Winfield[1]

Abstract

For many the Holy Grail of robotics and AI is the creation of artificial persons: artefacts with equivalent general competencies to humans. Such artefacts would literally be simulations of humans. With the theme of simulation this essay reflects on both the simulation of intelligence and the associated energy costs across three broad and controversial topics: first, how to design (or evolve) human-equivalent AI, second, the prospects of an intelligence explosion once the first has been achieved, and third, simulationism – the idea that we are ourselves simulations in a simulated universe.

Key words: artificial intelligence (AI), artificial general intelligence (AGI), energy, simulation, the singularity, simulationism.

1.1 Introduction

This essay offers some personal reflections on three broadly related subjects. The first is human-equivalent artificial intelligence (AI), that is an AI or a robot, its embodied counterpart, which would be regarded as having equivalent general competencies of a human (the question of what competencies and which human is one I will touch upon). The second, is the question of what happens after human-equivalent AI has been achieved: the technological singularity. And the third is simulationism: the idea that we are living in a computer simulation. These three subjects are linked in two ways. One is that they all deal with the simulation of complex systems – given that

an AI is a simulation of natural intelligence – and the other is energy, since simulation, like any physical process, costs energy.

1.2 Towards human-equivalent AI

1.2.1 Robots as smart-as-humans

In recent years we've seen many headlines confidently predicting robots with human equivalent AI. One such headline, from 2014, read '2029: the year when robots will have the power to outsmart their makers'[1], occasioned by an Observer interview with Google's director of engineering Ray Kurzweil. A survey of AI experts found a median prediction of a 50% chance of human-level AI during the 2040s [10].

Much as I respect Kurzweil's achievements as an inventor, I believe he is mistaken. Of course I can understand why he would like it to be so – he would like to live long enough to see this particular prediction come to pass. But optimism doesn't make for sound predictions. Here are several reasons that robots will, I believe, not be smarter than humans by 2029, 2039 or 2049.

- What exactly does as-smart-as-humans mean? Intelligence is very hard to pin down [1, 6]. One thing we do know about intelligence is that it is not one thing that humans or animals have more or less of [17]. Humans have several different kinds of intelligence – all of which combine to make us human. Analytical or logical intelligence of course – the sort that makes you good at IQ tests. But emotional intelligence is just as important, especially (and oddly) for decision making. So is theory of mind – the ability to infer others' beliefs and intentions.
- Human intelligence is embodied. As Pfeifer and Bongard explain in their outstanding book you can't have one without the other [11]. The old Cartesian dualism – the dogma that robot bodies (the hardware) and mind (the software) are distinct and separable – is wrong and deeply unhelpful. We now understand that the hardware and software have to be co-designed; animal brains and bodies clearly did not evolve independently. But we really don't understand how to do this – none of our engineering paradigms fit. A whole new approach may need to be invented.
- As-smart-as-humans probably doesn't mean as-smart-as newborn babies, or even two year old infants. Somehow-comparable-in-intelligence-to adult humans perhaps? But an awful lot happens between birth and adulthood. And the Kurzweilians probably also mean as-smart-as-very-well-educated-humans. But of course this requires both development – a

---

[1]   https://www.theguardian.com/technology/2014/feb/22/computers-cleverer-than-humans-15-years

lot of which somehow happens automatically – and a great deal of nurture. Again we are only just beginning to understand the problem, and developmental robotics – if you'll forgive the pun – is still in its infancy.

- Moore's Law will not help. Building human-equivalent robot intelligence needs far more than teraFLOPS of computing power. It will certainly need computing power, but that's not all. It's like saying that all you need to build a cathedral is 150,000 tons of marble. You certainly do need large quantities of marble – the raw material – but without (at least) two other things: the design for a cathedral, and the know how and tools to transform the architectural drawings into a building – there will be no cathedral. The same is true for human-equivalent robot intelligence.

- The hard problem of learning and the even harder problem of consciousness. (I'll concede that a robot as smart as a human doesn't have to be conscious – a philosophers-zombie-bot would do just fine.) But the human ability to learn, then generalise that learning and apply it to completely different problems is fundamental and remains an elusive goal for robotics and AI. This is called Artificial General Intelligence, which remains as controversial as it is unsolved.

These are the reasons I can be confident in asserting that robots will not be smarter than humans within 15 years. It's not just that building robots as smart as humans is a very hard problem. We have only recently started to understand how hard it is well enough to know that whole new theories (of intelligence, emergence, embodied cognition and development, for instance) will be needed, as well as new engineering paradigms. Even if we had solved these problems and a present day *Noonian Soong* had already built a robot with the potential for human equivalent intelligence – it still might not have enough time to develop adult-equivalent intelligence by 2029.

1.2.2 The energy cost of evolving human-equivalent AI

In the quest for human-equivalent AI there are, broadly speaking, three approaches open to us: design it, reverse-engineer it[2] or evolve it. The third of these – artificial evolution – is attractive because it sidesteps the troublesome problem of having to understand how human intelligence works. It's a black box approach: create the initial conditions then let the blind watchmaker of artificial evolution do the heavy lifting. This approach has some traction. For instance David Chalmers, in his philosophical analysis of the technological singularity [4], writes "if we produce an AI by artificial evolution, it is likely that soon after we will be able to improve the evolutionary algorithm and

---

[2] I will not discuss whole brain emulation (WBE) in any depth in this essay, but the WBE of a very simple animal – *C-elegans* – will play a part. For an account of WBE see Murray Shanahan's excellent book The Technological Singularity [12].

extend the evolutionary process, leading to AI+". And since we can already produce very simple AI by artificial evolution, then all that's needed is to "improve the evolutionary algorithm". If only it were that straightforward.

Some time ago I asked myself (and anyone else who would listen): ok, but even if we had the right algorithm, what would be the energy cost of artificially evolving human-equivalent AI? My hunch was that the energy cost would be colossal; so great perhaps as to rule out the evolutionary approach altogether. That thinking, and some research, resulted in a short paper in ALIFE 14 [16].

That paper explores the question: what is the energy cost of evolving complex artificial life? The paper takes an unconventional approach by first estimating lower and upper bounds on the energy cost of natural evolution and, in particular, the species *Homo Sapiens Sapiens*. A lower bound of $\sim 8000EJ^3$ is obtained by estimating and summing the calorific energy costs of populations of 7 concestor species stretching back from hominids to single-celled organisms over about 3.5B years. The upper bound of $\sim 5.7 \times 10^{12}EJ$ is obtained by simply estimating the amount of solar energy available to the evolution of all living things. Even though these are very crude estimates they have value because we are forced us to think about the energy costs of co-evolution, and hence the energy costs of evolving complexity.

Of course the processes and mechanisms of biological and artificial evolution are profoundly different (except for the meta-level equivalence of the Darwinian evolutionary operators: variation, selection and heredity), but there is an ineluctable truth: artificial evolution still has an energy cost. Virtual creatures, evolved in a virtual world, have a real energy cost. And we can estimate that energy cost. For a simple robot with an artificial neural network of comparable size to that of the nematode worm *C. elegans* (with 302 neurons), each simulated robot has a real energy cost of about $9J/hr$, which interestingly is about 2000 times greater than the energy cost of a very small (1mg) organism, $0.004J/hr$. It is clear that 'larger' artificial creatures, i.e. with more artificial neurons, must incur a greater computational energy cost.

In general, if the energy cost of simulating and fitness testing a virtual creature is e, then the energy cost of evolving that creature will be $E = gpe$, where g is the number of generations required and p the population size. Energy cost e is clearly a function of the complexity of that virtual creature, but how might e scale with complexity? Kleiber's law [5, p. 422] relates the mass of an organism to its energy consumption and, plotted on logarithmic axes, shows a remarkably consistent linear relationship from micro-organisms to the largest animals. Perhaps a similar relationship might exist between, say, neural complexity and energy cost e for virtual creatures: an artificial life equivalent of Kleiber's law?

Figure 1.1 imagines such a plot, of neural complexity against energy cost e. We cannot yet plot such a relationship since we have, to date, only one or

---

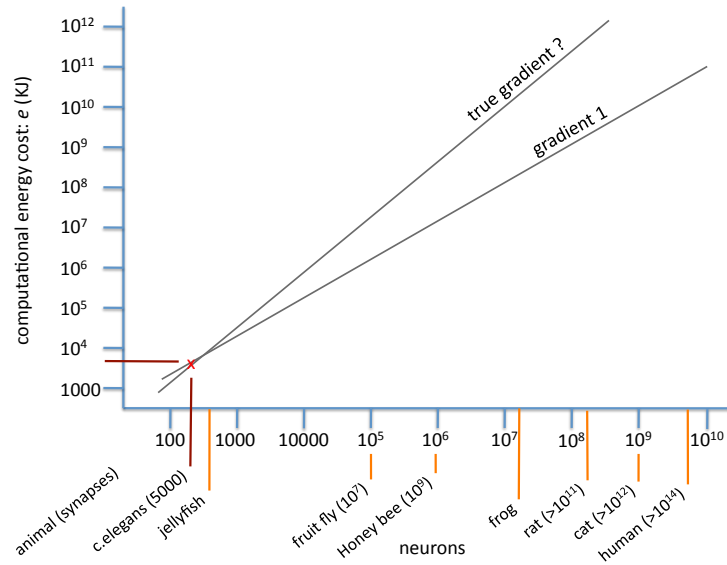^{3} An exajoule (EJ) is $10^{18}J$.

Figure 1. The computational energy cost of artificial neural complexity, after Kleiber's Law.

two points at the very bottom of the artificial neural complexity scale. But, if we assume that a human-equivalent AI will require roughly comparable neural complexity to *Homo Sapiens*[4], with 85×10^9 neurons and $10^{14} - 10^{15}$ synapses (and noting that neural complexity must take account of the number of synapses, since neural connections incur a computational energy cost), then e for an artificial creature of this synaptic complexity could be $10^{10} - 10^{12}$ times greater than for something equivalent to *C. elegans*. But this scale factor is still likely to be too low because fitness testing of increasingly complex artificial creatures will take longer and incur greater energy cost. It seems likely that the gradient of our ALife version of Kleiber's law will be greater than 1.

I estimate that the computational energy cost of simulating and fitness testing something with an artificial neural and synaptic complexity equivalent to humans could be around $10^{14} KJ$, or $0.1 EJ$. But evolution requires many generations and many individuals per generation, and many co-evolving artificial species. Also taking account of the fact that many evolutionary runs will fail (to produce smart AI), the whole process would almost certainly need to be re-run from scratch many times over. If multiplying those population sizes, generations, species and re-runs gives us (very optimistically) a factor of

---

[4] almost certainly not a safe assumption, but it's all we have to go on here.

1,000,000 - then the total energy cost would be $100,000EJ$. In 2010 total human energy use was about $539EJ$. So, artificially evolving human-equivalent AI would need the whole human energy generation output for about 200 years.

1.3 The Technological Singularity

The singularity. Or to give it it's proper title, the technological singularity. It's a Thing: an idea that has taken on a life of its own; more of a life, I suspect, than the very thing it predicts ever will. It's a Thing for the techno-utopians: wealthy middle-aged men who regard the singularity as their best chance of immortality. They are Singularitarians, some of whom appear prepared to go to extremes to stay alive for long enough to benefit from a benevolent super-AI – a manmade god that grants transcendence.

And it's a Thing for the doomsayers, the techno-dystopians. Apocalypsarians who are equally convinced that a superintelligent AI will have no interest in curing cancer or old age, or ending poverty, but will instead – malevolently or maybe just accidentally – bring about the end of human civilisation as we know it [3]. History and Hollywood are on their side. From the Golem to Frankenstein's monster, *Skynet* and *The Matrix*, we are fascinated by the old story: man plays god and then things go horribly wrong.

The singularity is basically the idea that as soon as artificial intelligence exceeds human intelligence then everything changes [12]. There are two central planks to the singularity hypothesis: one is the idea that as soon as we succeed in building AI as smart as humans then it rapidly re-invents itself to be even smarter, starting a chain reaction of smarter-AI inventing even-smarter-AI until even the smartest humans cannot possibly comprehend how the superintelligent AI works. The other is that the future of humanity becomes unpredictable and in some sense out-of-control from the moment of the singularity onwards.

So, should we be worried, or optimistic, about the technological singularity? Well I think we should be a little worried – cautious and prepared may be a better way of putting it – and at the same time a little optimistic (that's the part of me that would like to live in Iain M Banks' *The Culture*). But I don't believe we need to be obsessively worried by a hypothesised existential risk to humanity. Why? Because, for the risk to become real, a sequence of things all need to happen. Its a sequence of big **if**s. *If* (1) we succeed in building human equivalent AI and *if* (2) that AI acquires a full understanding of how it works[5], and *if* (3) it then succeeds in improving itself to produce super-intelligent AI, and *if* (4) that super-AI, either accidentally or maliciously,

---

[5] noting that we humans still have a *very* incomplete understanding of how we work.

starts to consume resources, and *if* (5) we fail to pull the plug then, yes, we may well have a problem. The risk, while not impossible, is improbable.

Each of these ifs needs detailed consideration. The first I have discussed in section 1.2 of this paper. Consider now the second: for that AGI to be able to understand itself well enough to be able to then re-invent itself – hence triggering an Intelligence Explosion is not a given. An AGI as smart and capable as most humans would not be sufficient – it would need to have the complete knowledge of its designer (or more likely the entire team who designed it) – and then some more: it would need to be capable of additional insights that somehow its team of human designers missed. Not impossible but surely very unlikely.

Take the third if: the AGI succeeds in improving itself. There seems to me no sound basis for arguing that it should be easy for an AGI – even one as smart as a very smart cognitive scientist – to figure out how to improve itself. Surely it is more logical to suppose that each incremental increase in intelligence will be harder than the last, thus acting as a brake on the self-improving AI.

By worrying unnecessarily I think we're falling into a trap: the fallacy of privileging the hypothesis. And – perhaps worse – taking our eyes off other risks that we should really be worrying about, like climate change, inequality or bioterrorism.

Wait a minute, I hear you say, there are lots of AI systems in the world already, surely it's just a matter of time? Yes we do have lots of AI systems, like chess programs, search engines, automated financial transaction systems, or the software in driverless car autopilots. And some AI systems are already smarter than all humans, like chess programs or AlphaGo[13]; some are smarter that most humans, like language translation systems. Others are as good as some humans, like driverless cars or natural speech recognition systems (like Siri) and sooner or later will be better than most humans. But none of this already-as-smart-as-some-humans AI has brought about the end of civilisation. The reason is that these are all narrow-AI systems: very good at doing just one thing.

A human-equivalent AI would need to be a generalist, like we humans. It would need to be able to learn, most likely by developing over the course of some years, then generalise what it has learned – in the same way that you and I learned as toddlers that wooden blocks could be stacked, banged together to make a noise, or as something to stand on to reach a bookshelf. It would need to understand meaning and context, be able to synthesise new knowledge, have intentionality and – in all likelihood – be self-aware, so it understands what it means to have agency in the world.

There is a huge gulf between present day narrow-AI systems and the kind of Artificial General Intelligence I have outlined. Opinions vary of course, but I think it's as wide a gulf as that between current space flight and practical faster than light spaceflight; wider perhaps, because we dont yet have a the-

ory of general intelligence, whereas there are several candidate FTL drives consistent with general relativity, like the Alcubierre drive[6].

So I don't think we need to be obsessing about the risk of superintelligent AI but I do think we need to be cautious and prepared. I strongly believe that all science and technology research should be undertaken within a framework of Responsible Innovation[9], and have argued that we should be thinking about subjecting robotics and AI research to ethical approval, in the same way that we do for human-subject research. Frameworks for the ethical design of robotics and AI have been developed in recent years, alongside professional codes of ethical conduct. IEEE Standards Association discussion document *Ethically Aligned Design* [7], for instance, has a section on the 'safety and beneficence of AGI and artificial superintelligence', outlining ethical issues and recommendations for addressing them[7].

1.4 Simulationism

Since Elon Musk's admission that he is a simulationist[8], the proposition that we are all living inside a computer simulation has attracted some attention. Of course the idea is not new. It can be traced back at least as far as Descartes' 'dream argument' and in recent years has been brought to prominence by Nick Bostrom's influential essay 'Are You Living In a Computer Simulation?'[2] My view is very firmly that the universe we are right now experiencing is real. Here are my reasons.

Firstly, Occam's razor; the principle of explanatory parsimony. The problem with the simulation argument is that it is a fantastically complicated explanation for the universe we experience. It's about as implausible as the idea that some omnipotent being created the universe. No. The simplest and most elegant explanation is that the universe we see and touch, both first hand and through our telescopes, LIGOs and Large Hadron Colliders, is the real universe and not an artifact of some massive computer simulation.

Second, is the problem of the reality gap[8]. Anyone who uses simulation as a tool to develop robots is well aware that robots which appear to work perfectly well in a simulated virtual world often don't work very well at all when the same design is tested in the real robot. This problem is especially acute when we are artificially evolving those robots. The reason for these problems is that the model of the real world and the robot(s) in it inside our simulation is an approximation. The reality gap refers to the less-than-perfect fidelity of the simulation; a better (higher fidelity) simulator would reduce the reality gap.

---

[6] https://en.wikipedia.org/wiki/Alcubierre_drive

[7] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_safety_beneficence_v2.pdf

[8] https://futurism.com/are-we-living-in-a-computer-simulation-elon-musk-thinks-so/

Anyone who has actually coded a simulator is painfully aware of the cost, not just computational but coding costs, of improving the fidelity of the simulation – even a little bit – is very high indeed. My long experience of both coding and using computer simulations teaches me that there is a law of diminishing returns, i.e. that the cost of each additional 1% of simulator fidelity is far more than 1%. I rather suspect that the computational and coding cost of a simulator with 100% fidelity is infinite. Rather as in HiFi audio, the amount of money you would need to spend to perfectly reproduce the sound of a Stradivarius ends up higher than the cost of hiring a real Strad and a world-class violinist to play it for you.

At this point the simulationists might argue that the simulation we are living in doesn't need to be perfect, just good enough. Good enough to do what exactly? To fool us that we're living in a simulation, or good enough to run on a finite computer (i.e. one that has finite computational power and runs at a finite speed). The problem with this argument is that every time we look deeper into the universe we see more: more galaxies, more sub-atomic particles, etc. In short we see more detail. The Voyager 1 spacecraft has left the Solar System without crashing, like *Truman*, into the edge of the simulation. There are no glitches like deja vu in *The Matrix*.

My third argument is about the computational effort, and therefore energy cost of simulation. I conjecture that to non-trivially simulate a complex system x (i.e. a human), requires more energy $e(x)$ than the real x consumes. An equation to express this inequality looks like this; how much greater depends on how high the fidelity of the simulation.

$$e(sim(x)) > e(x) \qquad (1.1)$$

Let me explain. The average human burns around 2000 Calories a day, or about 9000 KJoules of energy. How much energy would a computer simulation of a human require, capable of doing all the same stuff (even in a virtual world) that you can in your day? Well that's impossible to estimate because we can't simulate complete human brains (let alone the rest of a human). But here's one illustration. In 2016 Lee Sedol was famously defeated by AlphaGo [13]. In a single 2 hour match Sedol burned about 170 Calories – the amount of energy you'd get from an egg sandwich. In the same 2 hours the AlphaGo machine consumed around 50,000 times more energy.

What can we simulate? The only whole brain emulation of a complex organism demonstrated so far is the Nematode worm *C-elegans* – it is the only animal for which we have the complete connectome[9]. I estimate the energy cost of simulating the nervous system of a *C-elegans* is (optimistically) about 9 J/hour, which is about 2000 times greater than the real animal (0.004 J/hr).

What's more the relationship between energy cost and mass is logarithmic, following Kleiber's Law, and I strongly suspect the same law applies to scaling

---

[9] See the Open Worm project http://openworm.org/getting_started.html

up computational effort as outlined in section 1.2.2 above. If the complexity of an organism o is C, then following Kleiber's Law the energy cost of simulating that organism, e will be

$$e_o \propto C_o^{X(F)} \tag{1.2}$$

Furthermore, the exponent X (which in Kleiber's law is reckoned to be between 0.66 and 0.75 for animals and 1 for plants), will itself be a function of the fidelity of the simulation, hence $X(F)$, where F is a measure of fidelity.

By using the number of synapses as a proxy for complexity and making some guesses about the values of X and F we could probably estimate the energy cost of simulating all humans on the planet (very much harder would be estimating the energy cost of simulating every living thing on the planet). It would be a very big number indeed, but that's not really the point I'm making here.

The fundamental issue is this: if my conjecture that to simulate complex system x requires more energy than the real x consumes is correct, then to simulate the base level universe would require more energy than that universe contains – which is clearly impossible. Thus – even in principle – we could not simulate the whole of our own observable universe to a level of fidelity sufficient for our conscious experience. And, for the same reason, neither could our super advanced descendents create a simulation of a duplicate ancestor universe for us to (virtually) live in. Hence we are not living in such a simulation.

1.5 Some concluding remarks

For many researchers the Holy Grail of robotics and AI is the creation of artificial persons: artefacts with equivalent general competencies as humans. Such artefacts would literally be simulations of humans. Some researchers are motivated by the utility of AGI; others have an almost religious faith in the transhumanist promise of the technological singularity. Others, like myself, are driven only by scientific curiosity. Simulations of intelligence provide us with working models of (elements of) natural intelligence. As Richard Feynman famously said 'What I cannot create, I do not understand'. Used in this way simulations are like microscopes for the study of intelligence; they are scientific instruments [14].

Like all scientific instruments simulation needs to be used with great care; simulations need to be calibrated, validated and – most importantly – their limitations understood. Without that understanding any claims to new insights into the nature of intelligence – or for the quality and fidelity of an artificial intelligence as a model of some aspect of natural intelligence – should be regarded with suspicion.

In this essay I have critically reflected on some of the predictions for human-equivalent AI (AGI); the paths to AGI (and especially via artificial evolution); the technological singularity, and the idea that we are ourselves simulations in a simulated universe (simulationism). The quest for human-equivalent AI clearly faces many challenges. One (perhaps stating the obvious) is that it is a very hard problem.

However, I believe that the task is made even more difficult for two further reasons. The first is – as hinted above – that we have failed to recognize simulations of intelligence (which all AIs and robots are) as scientific instruments, which need to be designed, operated and results interpreted, with no less care than we would a particle collider or the Hubble telescope.

The second, and more general observation, is that we lack a general (mathematical) theory of intelligence. This lack of theory means that a significant proportion of AI research is not hypothesis driven, but incrementalist and ad-hoc. Of course such an approach can and is leading to interesting and (commercially) valuable advances in narrow AI. But without strong theoretical foundations, the grand challenge of human-equivalent AI seems rather like trying to build particle accelerators to understand the nature of matter, without the Standard Model of particle physics.

Acknowledgments

References

1. Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., et al (2018) Mapping intelligence: Requirements and possibilities. In: Müller, V. C., ed. (2018) Philosophy and Theory of Artificial Intelligence 2017 (PT-AI 2017). (44) Springer, Cham, pp. 117-135.
2. Bostrum, N. (2003) Are You Living In a Computer Simulation? Philosophical Quarterly, 2003, Vol. 53, No. 211, pp. 243-255.
3. Bostrom, N. (2014) Superintelligence: paths, dangers, strategies. Oxford University Press.
4. Chalmers, D. (2010). The singularity: A philosophical analysis. Journal of Consciousness Studies, 17 (9-10):765.
5. Dawkins, R. (2004). The Ancestors Tale. Weidenfeld and Nicolson, London.

6. Hernández-Orallo, J. (2017) The measure of all minds: evaluating natural and artificial intelligence, Cambridge University Press.
7. IEEE Standards Association (2017) Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, version 2, IEEE Standards Association. https://ethicsinaction.ieee.org/.
8. Jakobi, N., Husbands, P., and Harvey, I. (1995) Noise and the reality gap: the use of simulation in evolutionary robotics, in Advances in Artificial Life, 704-720, Springer Berlin.
9. Jirotka, M., Grimpe, B., Stahl, B., Eden, G., and Hartswood, M. (2017) Responsible research and innovation in the digital age. Commun. ACM 60, 6268.
10. Müller V.C. and Bostrom N. (2016) Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In: Müller, V. C., ed. (2016) Fundamental Issues of Artificial Intelligence. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), vol 376. Springer, Cham
11. Pfiefer, R. and Bongard, J. (2007) How the body shapes the way we think: a new view of intelligence. MIT Press.
12. Shanahan, M. (2015) The Technological Singularity, MIT Press.
13. Silver, D., et al. (2016). Mastering the game of go with deep neural networks and tree search. Nature, 529, 484489.
14. Stepney, S., Polack, F.A.C., Alden, K, Andrews, P.S. et al (2018) Engineering Simulations as Scientific Instruments: A Pattern Language, Springer.
15. Walsh, R. and Stepney, S. (eds) (2018) Narrating Complexity, Springer.
16. Winfield, A.F. (2014) Estimating the energy cost of (artificial) evolution. In: Sayama, H., Rieffel, J., Risi, S., Doursat, Ren and Lipson, H., eds. Artificial Life 14: 14th International Conference on the Synthesis and Simulation of Living Systems, New York, 31 July - 2 August 2014., pp. 872-875
17. Winfield, A.F. (2017) How intelligent is your intelligent robot? arXiv preprint arXiv:1712.08878