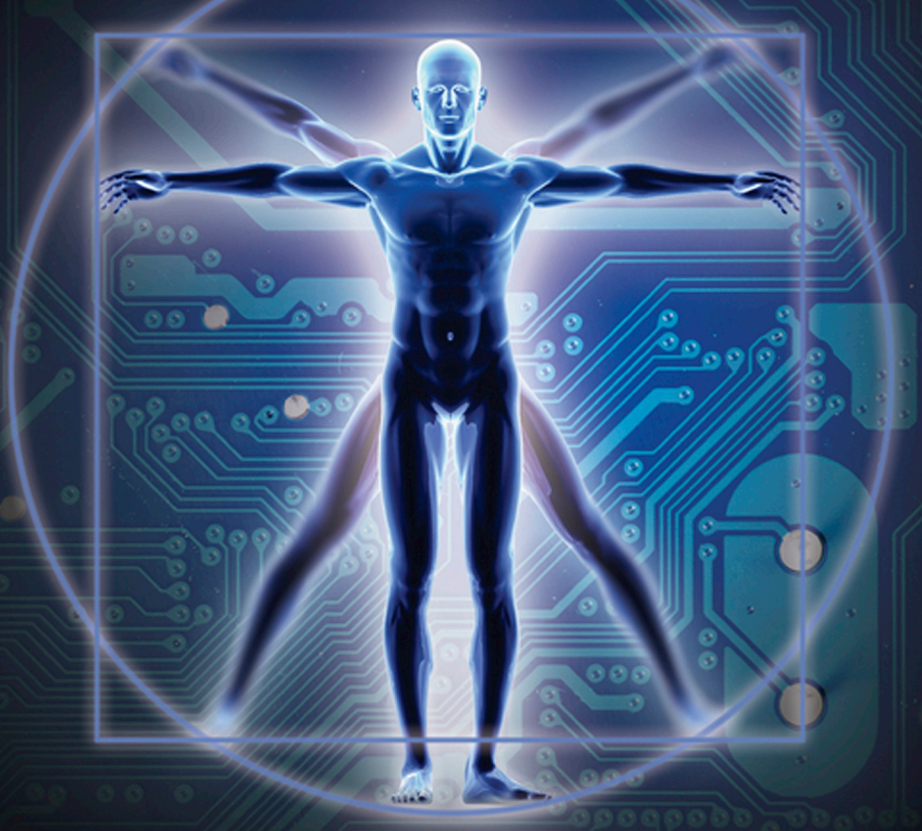
Intelligence Unbound

The Future of Uploaded and Machine Minds

Edited by

Russell Blackford and
Damien Broderick



WILEY Blackwell

Intelligence Unbound

Intelligence Unbound

The Future of Uploaded and Machine Minds

Edited by

Russell Blackford and
Damien Broderick

WILEY Blackwell

This edition first published 2014

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data applied for.

Hardback ISBN: 978-1-118-73641-8
Paperback ISBN: 978-1-118-73628-9

A catalogue record for this book is available from the British Library.

Cover image: Circuit board © sborisov /iStockphoto; Vitruvian Man © Devrimb /iStockphoto.

Set in 10/12pt Sabon by Laserwords Private Limited, Chennai, India

1 2014

To Aubrey Townsend, who handed me the tools
Russell Blackford

To R. Daneel Olivaw, Golem XIV, Donovan's Brain, and Paul Durham, in the hope that things turn out better next time
Damien Broderick

Contents

Notes on Contributors

Nicholas Agar is a New Zealand philosopher, based at Victoria University of Wellington. His research is focused on ethical issues arising out of the application of new technologies to human beings. His most recent books are *Humanity's End: Why We Should Reject Radical Enhancement* (2010) and *Truly Human Enhancement: A Philosophical Defense of Limits* (2014).

Michael Anissimov is a futurist focused on such emerging technologies as nanotechnology, biotechnology, robotics, and artificial intelligence. He previously managed the Singularity Summit and worked as media director for the Machine Intelligence Research Institute, as well as co-founding Extreme Futurist Festival.

Stuart Armstrong and Seán ÓhÉigeartaigh work at the Future of Humanity Institute of Oxford University, where they analyze the major risks facing humanity, and how these can be prevented or mitigated. Recent work has focused on the risks and ethics of AI, human biases, and the reliability of predictions.

Russell Blackford is an Australian philosopher and literary critic. He is a Conjoint Lecturer at the University of Newcastle, NSW, and editor-in-chief of *The Journal of Evolution and Technology*. His recent books include *Freedom of Religion and the Secular State* (2012) and *Humanity Enhanced: Genetic Choice and the Challenge for Liberal Democracies* (2014).

James Bodington is a doctoral candidate in philosophy at the University of New Mexico, where he studies twentieth-century and contemporary continental philosophy, especially philosophy of religion and philosophy of

technology, with a particular emphasis on the ethical and political ramifications of technology.

Damien Broderick holds a PhD in the literary theory of the sciences and the arts from Deakin University, and has written or edited some 60 books in several disciplines, including a number of prize-winning novels. His *The Spike* (1997, 2001) was the first general treatment of the Singularity. In 2008 he edited an original science anthology *Year Million*, on the prospects of humankind in the remote future.

David J. Chalmers is Distinguished Professor of Philosophy and Director of the Centre for Consciousness at the Australian National University and Professor of Philosophy at New York University. He is best known for articulating what he has dubbed the "hard problem" of consciousness – explaining how physical brains and bodies give rise to "qualia," or subjective experiences. His best-known book is *The Conscious Mind* (1996).

Joseph Corabi is an Associate Professor of Philosophy at Saint Joseph's University. He has published numerous articles on philosophy of mind and philosophy of religion.

Linda MacDonald Glenn is a bioethicist, healthcare educator, lecturer, consultant, and attorney. She holds faculty appointments at the Alden March Bioethics Center and California State University Monterey Bay. Her research is focused on the sociopolitical implications of exponential technologies and evolving concepts of legal personhood.

Ben Goertzel, PhD, chief force behind the recent movement toward artificial general intelligence in the AI field, is chief scientist of financial prediction firm Aidyia Holdings and chairman of AI software company Novamente LLC and bioinformatics company Biomind LLC. His research work encompasses artificial general intelligence, natural-language processing, cognitive science, data mining, machine learning, computational finance, bioinformatics, virtual worlds, gaming, and other areas.

Kathleen Ann Goonan is the author of *Queen City Jazz* (1994), *The Bones of Time* (1996), *Mississippi Blues* (1998), *Crescent City Rhapsody* (2000), *Light Music* (2002), *In War Times* (2007), *This Shared Dream* (2011), and *Angels and You Dogs* (2012). She is a Professor of the Practice at Georgia Institute of Technology, Atlanta, where she teaches creative writing and examines the intersection of culture, science, technology, and literature. Her website is www.goonan.com.

Victor Grech is Consultant Pediatrician (Cardiology), Pediatric Department, Mater Dei Hospital, Tal-Qroqq, Malta, and author of several searching essays on the thematics of *Star Trek*.

Robin Hanson is an Associate Professor of Economics at George Mason University and a research associate at the Future of Humanity Institute of Oxford University. He is known as an expert on prediction markets and was a principal architect of the Foresight Exchange, DARPA's FutureMAP project, and IARPA's DAGGRE project.

James J. Hughes is the Executive Director of the Institute for Ethics and Emerging Technologies, and a bioethicist and sociologist at Trinity College in Hartford, Connecticut, where he teaches health policy. Hughes is author of *Citizen Cyborg*, and is working on a second book tentatively titled *Cyborg Buddha*.

Randal A. Koene introduced the multidisciplinary field of whole brain emulation and is lead curator of its scientific roadmap. He is founder of the Carboncopies.org foundation and neural interfaces company NeuraLink Co, and Science Director of the 2045 Initiative. His publications, presentations and interviews are available at http://randalkoene.com.

Richard Loosemore is a lecturer in the Department of Mathematical and Physical Sciences at Wells College. He graduated from University College London as a physicist and from Warwick University as a cognitive scientist. His background includes work in artificial intelligence, cognitive science, physics, software development, philosophy, parapsychology, and archeology.

Max More, who received his PhD in philosophy from the University of Southern California, is a strategic philosopher recognized for his thinking on the implications of emerging technologies. More's contributions include founding the philosophy of transhumanism, developing the Proactionary Principle, and co-founding Extropy Institute. He is currently president and CEO of the Alcor Life Extension Foundation.

Seán ÓhÉigeartaigh and Stuart Armstrong work at the Future of Humanity Institute of Oxford University, where they analyze the major risks facing humanity, and how these can be prevented or mitigated. Recent work has focused on the risks and ethics of AI, human biases, and the reliability of predictions.

Nicole Olson is a Canadian transhumanist writer/researcher holding a bachelor's degree from the University of Alberta in philosophy and sociology.

Massimo Pigliucci is a Professor of Philosophy at the City University of New York. His research is concerned with philosophy of science, the relationship between science and philosophy, and the nature of pseudoscience. His publications include several books, most recently

Answers for Aristotle (2012) and *Philosophy of Pseudoscience* (co-edited with Maarten Boudry, 2013).

Joel Pitt, PhD, is a scientist and software developer based in Wellington, New Zealand. As a scientist, he has contributed original research to molecular biology, machine learning, and ecology. As a developer he has been the CTO Demand Analytics, and currently works for Dragonfly Data Science, a science consultancy in Wellington.

Anders Sandberg has a background in computational neuroscience and the ethics of human enhancement. Since 2008 he has been James Martin Research Fellow at the Future of Humanity Institute at Oxford University, where he is investigating neuroethics, global catastrophic risks, and applied epistemology.

Susan Schneider is an Associate Professor of Philosophy at the University of Connecticut. She has published many articles in the fields of metaphysics and philosophy of mind as well as *The Language of Thought, The Blackwell Companion to Consciousness* (with Max Velmans), and *Science Fiction and Philosophy*.

Joe Strout's career blends science and technology, with degrees in psychology and neuroscience, and extensive experience as a software engineer. He works now as a software consultant, developing artificial intelligence algorithms for the game industry, as well as other applications in business and medicine. His website is http://www.ibiblio.org/jstrout/uploading.

Iain Thomson is Professor of Philosophy at the University of New Mexico. The author of two books, *Heidegger on Ontotheology: Technology and the Politics of Education* (2005) and *Heidegger, Art, and Postmodernity* (2011), Thomson has published dozens of articles in philosophical journals, essay collections, and reference works, and his writing has been translated into seven languages.

Natasha Vita-More, PhD, is a Professor at the University of Advancing Technology and Founder of H+ Lab. She has appeared in over 24 televised documentaries and featured in *Wired*, the *New York Times*, and *Village Voice*. She is chair of Humanity+ and a Fellow of the Institute for Ethics and Emerging Technologies.

Mark Walker is an Associate Professor in the Department of Philosophy at New Mexico State University, where he holds the Richard L. Hedden Chair of Advanced Philosophical Studies. His book, *Happy-People-Pills for All* (2013) argues for creating advanced pharmaceuticals to boost the happiness of the general population.

Naomi Wellington is a postgraduate philosophy student at the Australian National University, working under the supervision of Daniel Stoljar and David Chalmers. Her academic background includes a BA with philosophy honors (H1) from Monash University. Her primary areas of interest are philosophy of mind and philosophy of neuroscience.

Introduction I: Machines of Loving Grace (Let's Hope)

Damien Broderick

1 Machine minds or humans copied into machines?

In an immensely confident but typical summary of the neurocomputational model of mind now dominant in science, Nobel Laureate Eric Kandel wrote in 2013:

> This new science of mind is based on the principle that our mind and our brain are inseparable. The brain is a complex biological organ possessing immense computational capability: it constructs our sensory experience, regulates our thoughts and emotions, and controls our actions. It is responsible not only for relatively simple motor behaviors like running and eating, but also for complex acts that we consider quintessentially human, like thinking, speaking and creating works of art. Looked at from this perspective, our mind is a set of operations carried out by our brain.[1]

More than two decades earlier, the science fiction writer Charles Platt offered a somewhat ampler view:

> A person's mind is structure as well as content. Without the structure, the content can't function. Our minds have to have the specialized architecture … in which to operate. We can store our brain data elsewhere, but when we do that, it's as nonfunctional as a videodisc without a disc player. (Platt 1991: 238)

In the next 25 to 100 years, genuinely intelligent machines are likely to be developed up to and beyond the highest levels of human ability.

Intelligence Unbound: The Future of Uploaded and Machine Minds, First Edition.
Edited by Russell Blackford and Damien Broderick.

We're not there yet, in part because the raw computational power of the brain hugely outstrips even the fastest computer. In mid-2013, the world's top supercomputer was the Tianhe-2, holding more than a million gigabytes of memory and running at some 50 petaflops (where a petaflop is a thousand trillion calculations per second) on its best days. Using an only slightly less extraordinary machine, Japan's 10 petaflop K supercomputer, scientists simulated 1 percent of 1 second of human brain activity. That took 40 minutes of screamingly fast calculations.[2]

You would need to multiply that by a factor of a quarter million to emulate a brain. Luckily, Moore's law (roughly: "computer power doubles every year and a half"[3]) suggests that a machine of this majestic status will be available – all things going well – in perhaps 30 more years. And of course in the meantime, scientists might learn better ways to get the job done sooner on leaner computers.

> Markus Diesmann of the Institute of Neuroscience and Medicine at Germany's Forschungszentrum Julich believes that, within the next decade, we'll be able to use exascale computers – capable of 1000 times one quadrillion operations per second – to represent the entire [*sic*] of the brain "at the level of the individual nerve cell and its synapses."[4]

And Henry Markram, a professor of neuroscience at the Swiss Federal Institute for Technology and founder and director of the Blue Brain Project, is coordinating the Human Brain Project (Keats 2013). This 10-year, €1.3 billion flagship project, selected in January 2013 by the European Commission, plans to simulate

> a rat cortical column. This neuronal network, the size of a pinhead, recurs repeatedly in the cortex. A rat's brain has about 100,000 columns of [about] 10,000 neurons each. [A] human cortex may have as many as two million columns, each having [about] 100,000 neurons each … These models will be basic building blocks for larger scale models leading towards a complete virtual brain.[5]

Will they necessarily be conscious, such brainy machines? Perhaps not, or at any rate not as we experience consciousness. The speeding locomotive, or "Iron Horse," never resembled a real horse, yet it carried a heavier load and moved much more swiftly and without tiring. A submarine isn't much like a whale, yet dives deeper and travels faster. Birds sing more beautifully than jet planes or rockets, but their capacity to fly high, far, and rapidly was outstripped by machines a century ago. Chess programs defeat grand masters without being self-aware, and IBM's Watson supercomputer beat top human contestants on *Jeopardy!*, winning a million dollars but without a jitter of anxiety or a shout of joyful pride.

Even so, machine or artificial intelligence (AI), unlike ours, might well have the ability to understand, modify, and improve its own source code, carrying it by great leaps into domains of ability that unaided flesh can never hope to reach. Half a century ago, the mathematician I.J. Good proposed that an "ultraintelligent machine" could design ever more enhanced versions of itself, resulting in an "intelligence explosion" that would leave humans far behind (Good 1965). If such supersmart computers also achieve consciousness, we (or our children and grandchildren) shall share the planet with a new and intriguing species of mentality.

But wait – what *is* intelligence? Thousands of learned books and scientific or philosophical papers have probed every corner of this apparently simple question with no clear consensus emerging. We can start with theoretical neurophysiologist William Calvin's breezy summary in *How Brains Think*:

> I think of intelligence as the high-end scenery of neuro-physiology – the outcome of many aspects of an individual's brain organization which bear on doing something one has never done before … some of *what* intelligence encompasses are cleverness, foresight, speed, creativity, and how many things you can juggle at once. (1997: 11)

Instead of our brutally slow chemical neurotransmitters, ionic currents, and neural designs, built by millions of years of ad hoc evolution, AI will use engineered electronic or photonic neural nets operating a million times faster. Instead of memories limited by the gene-architected size of our skulls and the human birth canal, AIs will possess effectively limitless storage constrained only by pathways traversed at the speed of light. In that sense, the arrival of advanced AIs will mark the end of some of the limitations that bind human intelligence. Intelligent and superintelligent machines will truly represent "intelligence unbound."

If and when this happens, humanity will face ethical issues of unprecedented gravity and difficulty. What obligations do we owe to artificial minds? Can they morally be switched off, like any other instrument or mechanical device? Or do they share human rights to life and the pursuit of happiness, the right of due process? Is there any way in which their designers can defang hazardous AIs that might turn on us, can make them compliant, obedient to their creators? Or is that slavery, mind bondage? If they are our intellectual superiors, can they at least be encouraged to adopt an attitude of benevolence toward us? Is it even technically possible to enforce friendship between protein and silicon beings, once the AIs pass beyond human comprehension in their abilities and potential?

In addition to this vexed and giddy outlook, in a near future of such fabulous machines, it will be possible to blend human and machine by enhancing

our current bodies with chips, modules, and interface devices (a process of "cyborgization" that has already begun).

All these prospects, and more, are discussed in detail in the chapters of this book. No single viewpoint is privileged throughout; these topics remain genuinely controversial, even philosophically troubling, so it is necessary to approach the topics carefully, exploring the pros and cons. And if the imminent arrival of machines with intelligence, however alien, is sure to throw our world into confusion and tumult, how much more will the possibility of minds copied from organic brains to inorganic machines? Not just copied as a static representation, as the Mona Lisa might be counterfeited with great fidelity by a skilled artist, but imbued with emotion, awareness, and all the other aspects of personhood.

2 Emulating the mind

This radical option might become available alongside the emergence of machines powerful enough and intricately connected enough to house a true mind. In the process – called "uploading" by some and, confusingly, "downloading" by others, and "whole brain emulation" by a third group – we could *become* machines while remaining ourselves, physically transferring the structure of our minds into capacious computer programs that generate thought and the quality of minds when they are run. Uploads would live in vivid virtual realities fitted to the needs of their simulated minds, while remaining in touch with the external world.

Is this a crypto-religious hope, the much-lampooned "Rapture of the Nerds"? It does echo religious hopes of reincarnation widespread in Asian cultures, where a non-material essence slips out of an injured or aged body to enter the waiting vessel of an unborn infant. But no, the prospect of uploading has nothing significantly in common with those ancient wishful, consoling dogmas. Naturalistic materialism, the current scientific paradigm, maintains that mind is nothing other than the sublimely complex workings of the physical brain and its bodily extensions in a world of particles and force fields. If that is what we *are*, nothing prevents us from copying – mapping – our neurological complexity into some more durable, swifter material substrate.

Still, isn't this a version of the cliché from bad horror movies: a naked brain in a vat of chemical soup? Some will complain that uploading is a nightmare proposed by body-hating, frightened computer hackers, those nerdish social incompetents allegedly fleeing from sensuous reality and human warmth.

It is true that many proponents of uploading dislike the limitations and messy urgings of the body and its ancient, now often maladaptive Darwinian drives. For others, as Max More details in his chapter, what drives the interest in uploading is a desire for more life, for the greatest possible access to this beautiful and complex universe. It can't be explained away as simple hatred or fear of the flesh.

Suppose it is true that mind and passion and soul are indeed the body at work, a whirling composite of matter and force and energy, engaged with the world. As we eat, drink, and excrete, the very atoms in our cells are regularly replaced. Should we object if mind changes its location from one kind of organized and ceaselessly replaced matter to another material substrate?

It is easy to become trapped by old preconceptions. Is the mind really a machine? If being a machine suggests clockwork or even the relatively stupid computers in our smart phones (already 50 percent more powerful than the greatest supercomputers in 1976), of course not. Even these limited computers are vastly more complex than an eighteenth-century wind-up parrot, or a nineteenth-century piano driven by a paper tape. The human brain is not like a broken-down motor-mower, and nobody ever thought it was.

Uploading need not imply a world of bloated grubs lying in the dark with their brains wired to spreadsheets and simulated worlds. On the contrary: transhumanist philosopher Max More, who intends to upload when that becomes an option (and use his new freedom to explore the stars), put his own case back in the 1990s: "I'm in the gym five days a week, plus I either run or cycle. I can boast that I do 710 lbs on the leg press. No atrophied body here!" In 2013, he added with amusement, "That was 710 lbs for 8 repetitions. I'm currently doing 720 lbs for 15 reps, so I'm definitely stronger. For 8 reps, I can do something over 800 lbs" (private communication). The initial goal of uploaders would be to emulate and enhance the brain, and that requires rich connections to external reality. It calls for give and take, building from the peculiar truth that inside our porridge-like brain matter is where our selves are generated. That fact does not repudiate the body, far from it.

A quadriplegic with no access to the world other than her mouth and ears and eyes and her vivid, courageous brain *is a person*. By contrast, the superb corpse of an Olympic athlete or concert pianist with a fatal brain injury, its metabolism sustained by medical machines, is no kind of person at all, just a tragic reminder of the fallibility of life and a storehouse for luckier transplant patients.

It's worth noting that if synthetic neurons can be made half the size of the organic varieties, replacing each brain cell after copying its structure

and contents, you could *double* the number of neurons inside your head. Would this automatically increase your brain power? Perhaps not, because specialized architecture is crucial to cognition. Still, one of the most palpable differences between modern humans and Lucy, the proto-hominid of the Ethiopian plains 3 million years ago, is that we have, on average, 1330 grams of brain tissue, while she had to make do with a third of that. With more components and some measure of plasticity in rewiring them, we might find ourselves becoming conspicuously cleverer in the ensuing weeks and years.

Suppose, once the mapping and substitution are complete, that this detailed atlas of your brain is also copied into the huge memory and processing system of a supercomputer – one easily handling as many petaflops or even exaflops as a human brain – perhaps a hundred thousand trillion calculations each second. So this mindless, terribly fast machine now contains a digital description that emulates your original brain. If we arrange for streams of data from the outside world to enter its ports, just the same kinds that now enter your own ears and eyes and taste buds and movement sensors and internal monitors scattered through your organic body – what happens? Add efferent (outgoing) channels that permit the emulated brain to reach into the world, to hold and move objects, stroke and sniff and chew in synchrony with the afferent (incoming) impulses that feed its sensorium. The simulated "you" will then, surely, feel himself or herself to "be" a "person" – to be, in fact, *you*!

Assuming we've left nothing out in this exercise in simulation, what we've achieved in our thought experiment is indeed an *upload* – a complete copy of your mind into the flickering electronic whir of a computer platform. Just to make sure you don't go mad at the shock of the transition, the flow of sensory data will ensure that you genuinely *feel* a physical continuity with your old self.

This might require linking your cybermind to a humanoid robot extension replete with stereo TV cameras at the top, and two servo-mechanical arms hinged at elbow and shoulder, and five tactile, gripping fingers, and two prosthetic legs. Alternatively, your experience might be delivered, after considerable pre-processing, in the form of a convincing virtual reality construct. Only that small part of the world you're choosing to look at will exist in the model, but that portion will be rendered with the maximum available pixel-rich detail at the focal zone, fading away to impressions at the boundaries – just like now, in fact.

Isn't this the worst kind of manipulative dehumanization ever proposed? Consider the nauseated response likely from humanists such as the American novelist Jonathan Franzen, who recently deplored how Twitter and Facebook, in his view, are diminishing the richness of human life. He gestures at an

> apocalyptic argument about the logic of the machine, which has now gone global and is accelerating the denaturization of the planet and sterilization of its oceans. I could point to the transformation of Canada's boreal forest into a toxic lake of tar-sands byproducts, the leveling of Asia's remaining forests for Chinese-made ultra-low-cost porch furniture at Home Depot, the damming of the Amazon and the endgame clear-cutting of its forests for beef and mineral production, the whole mindset of "Screw the consequences, we want to buy a lot of crap and we want to buy it cheap, with overnight free shipping." (Franzen 2013)

And meanwhile, some scientists and futurists propose whole brain emulation or uploading into the technoapocalyptic machine? For many, this will seem the ultimate blasphemy.

3 Is my copy me?

Here, then, are the key questions:

- Is this uploaded personality conscious?
- Is it (he, she) you or your twin, or something unprecedented?
- If a disaster destroys the hardware it's running on, and the latest back-up is reinstalled on a new machine, is the new version of you the same as the original? After all, you remain *you* when you sleep and wake. Or is it quite a different person, who just happens to recall everything that ever happened to you (with at least the same fidelity now available to your own organic brain)?
- Would you even be prepared to terminate your own stream of awareness, just so that this other person (with the same memories, admittedly) could awaken to an adventure in virtual reality, to potentially endless machine life?

Some say yes. After all, once you're inside that computer, the benefits are extraordinary. No more colds or cancer. Your thinking might accelerate. No longer restricted to the sluggish baton-passing of neurotransmitters and ionic currents, your electronic or optical consciousness-stream would blaze like the pure spirit of some Miltonic angel. And those convenient backups provide a degree of security the flesh can never maintain.

But wait – why should you care about *his* or *her* greater wealth? *Their* security of tenure in a perilous universe?

Evolution has winnowed our genomes in favor of building bodies whose economic behavior is channeled by the need to sustain (and reproduce) the components of that genotype – the individual genes, and the ensembles of genes that do well together.

According to a standard argument, we tend to be "altruistic" toward other bearers of large chunks of the same genotype. Note one crucial and somewhat paradoxical proviso: this mechanism allows high-level adapted structures such as brains and cultures to make "mistaken" identifications. Individuals can sacrifice themselves in support of genotypes quite different from their own. Young men hurl themselves into the firing line because, in a sense, their genetic propensities have been tricked, persuaded to bond with their (genetically distant) fellow warriors, or with their nation or religion. Still, it's clear that the altruism equations would be more than satisfied if you were to sacrifice your body in order to produce a dozen copies of your exact genome, with or without cultural and individual memories.

You the original might not be so ardent.

Grant that you could arrange for a dozen exact copies of your body by cloning: a time-lapsed set of identical -tuples. If you could only achieve this by giving up your individual life you might not share the joy.

Suppose, however, that these copies could also contain your exact memories to this moment, as we've stipulated above, so that you spawned a dozen true copies of yourself. (Admittedly, these duplicates would fork off from each other immediately in terms of experience.) Would this offer make you more inclined to die in order to achieve perfect duplication? Possibly not. Why should you care about *their* enhanced prospects, even if each one of them is convinced that he or she *is* you?

But suppose you were dying of a currently incurable disorder, and you are offered a new, young replacement body grown from your stem cells and imprinted with all your memories, or a robot equivalent, or an emulated and superior cyber version inside an exaflop machine. Would you prefer to go instead gently and immediately into that good night?

These are *difficult* issues, and the answers are not at all self-evident.

Indeed, you might not even be at immediate risk of dying, except in the sense that we all are. Perhaps lingering death by senility or disease causes the irreparable loss of too many brain cells to permit machine resurrection. Perhaps the sooner you get your brain destructively scanned the more reliable the result will be. What would *you* do?

4 Who woke up?

Physical and general brain-process continuity does seem to support our current sense of continuing identity. This allows us to believe that the person who wakes up tomorrow is the same one who went to sleep tonight. What of awakening after 10 years in a coma? Nobody denies that this case is deeply traumatic, but still we assume that *the same person* has woken up. Waking

after half the brain is removed to forestall death by cancer? No one denies that this case is even more deeply troubling.

Would you be prepared to die (sacrifice your current embodiment) in order that an exact copy of yourself be reconstituted elsewhere (teleported), or on a different substrate? To insist that this question remains unsettled (and unsettling!) is not to be a hostile "upload skeptic."

In a sense, the philosophical question is moot. As long as there are people who share this conviction that identity persists through upload, they'll purchase the service the moment it is technically feasible. Unless their machine emulations regret the error of their choice (and why should they, since they share the memories and disposition of their predecessors?),[6] humankind will be split into two species: people in their original and continuing bodies, and uploaded people who are copies of dead humans. Add to that a third category, independent minds built from the ground up by artificial intelligence programs. And perhaps a fourth: living, embodied humans in an extended condition of enhancement, plugged in (perhaps only some of the time) to the super-Net.

What happens in a world like this? How do we deal with such a proliferation of new intelligent species? Even before superintelligence makes the world more fraught for ordinary people than it is now – and this could occur due to genetic engineering, as well as AI or uploading – we will find ourselves in a disrupted psychic ecology. Species competition is fiercest between relatives. We do not usually fight with birds and bees for living space, not as the military of nations do. The soil continues to be churned by uncaring worms, whoever walks upon its surface. Our ancient ancestors, admittedly, did wipe out all the other hominins and primates that got anywhere near our ecological zone. Of our omnivorous cousins, only the chimpanzees and bonobos remain, and they persist at our sufferance. True AIs and uploads will suffer the same Darwinian neighborhood pressures, as will we.

At last, even if your original brain is altogether gone, your mind (or its perfect copy) remains as active as ever. Once we command the new array of expanded senses, the unprecedented range of access open to us, we would less *inhabit* our new locus of consciousness than *be* it. And strange new doorways will open to futures with entirely new ethical quandaries, hazards, and joys that are barely imaginable.

Notes

1 Kandel shared the 2000 Nobel Prize in Physiology or Medicine, and is a professor at the Mortimer B. Zuckerman Mind Brain Behavior Institute at Columbia, and a senior investigator at the Howard Hughes Medical Institute. He is an editor

and author of *Principles of Neural Science*, 5th edn. (New York: McGraw-Hill, 2013).

2 http://www.top500.org/blog/lists/2013/06/press-release/ (accessed October 6, 2013).

3 Precisely, the "law" is a historical observation that the number of transistors on an integrated circuit chip doubles every two years, while performance increases by a factor of two every 1.5 years or even faster. It has held true since 1958.

4 http://www.top500.org/blog/an-83000-processor-supercomputer-can-only-match-1-of-your-brain/ (accessed September 12, 2013).

5 Blue Brain Project, http://bluebrain.epfl.ch/page-56882-en.html

6 They do, however, in Norman Spinrad's interesting short novel *Deus X* (1993), which hangs on this very point. And in Greg Egan's brilliant upload novel *Permutation City* (1994), uploads have a distressing way of killing themselves the moment they understand that they are the copies in a virtual reality world and not the originals. See also Platt 1991.

References

Calvin, William H. 1997. *How Brains Think: Evolving Intelligence, Then and Now*. London: Weidenfeld & Nicolson.

Egan, Greg. 1994. *Permutation City*. London: Millennium.

Franzen, Jonathan. 2013. What's wrong with the modern world. *The Guardian*, http://www.theguardian.com/books/2013/sep/13/jonathan-franzen-wrong-modern-world (accessed October 7, 2013).

Good, I.J. 1965. Speculations concerning the first ultraintelligent machine. *Advances in Computers*, 6. New York: Academic Press.

Kandel, Eric R. 2013. The new science of mind. *New York Times*, http://www.nytimes.com/2013/09/08/opinion/sunday/the-new-science-of-mind.html (accessed October 7, 2013).

Keats, Jonathon. 2013. The $1.3B quest to build a supercomputer replica of a human brain. *Wired*, http://www.wired.com/wiredscience/2013/05/neurologist-markam-human-brain/all/ (accessed 7 October 7, 2013).

Platt, Charles. 1991. *The Silicon Man*. New York: Bantam Books.

Spinrad, Norman. 1993. *Deus X*. New York: Bantam Spectra.

Introduction II: Bring on the Machines

Russell Blackford

1 A strange new epoch?

Year by year, computer programs and applications become more fluid, more dazzling, more convenient for our purposes. The allure is obvious, and the results are often beneficial, but how far can we go with the rise and rise of digital technology? We've come very far already, if we're counting sheer numbers:

> [B]etween 2008 and 2009 … the number of devices – sensors, phones, computers – connected to the Internet outnumbered the human population. By 2020, when an estimated 7.6 billion people will be running around on the planet, there will be 50 billion machines communicating to one another. (Tucker 2013)

Clever mobile apps can already help with many of our daily problems, even, as James Hughes' chapter shows in detail, providing us with moral support and guidance. While there may be little prospect of machines and apps answering the deepest moral questions, they can remind and exhort us about many of our goals and the steps prescribed to achieve them. As Hughes makes clear, our mobile devices are becoming tutors, counselors, and externalized consciences.

So far, so good, but more radical outcomes may not be far away. To examine the implications, Damien Broderick and I obtained chapters from a brilliant, clued-up, and diverse cast of contributors. They bring expertise from

Intelligence Unbound: The Future of Uploaded and Machine Minds, First Edition.
Edited by Russell Blackford and Damien Broderick.

the fields of philosophy and cognitive science, from art and literature, and from the burgeoning young culture of transhumanism. The book you're reading, perhaps on your computer or an e-book reader or even the tiny screen of a smartphone, deals with the prospects of full-scale machine intelligence and what has become known as mind uploading – the idea that living, functioning minds could be transferred to powerful computer hardware.

If we take Hans Moravec, the celebrated Carnegie Mellon robotics guru, as our guide to the future, a strange epoch is approaching. In his 1988 book *Mind Children*, Moravec argues that culture and intelligence will soon be the province of increasingly capable machines. He dares us to contemplate what we cannot genuinely imagine: the post-biological world of our thinking, ever-improving "unfettered mind children" (Moravec 1988: 5), superintelligent machines that aspire to immortality and endless self-replication. In this vision, flesh-and-blood human bodies will "play a rapidly diminishing role" (1988: 4). The machines will supersede us, or perhaps they will *be* us – greatly transformed by technology – since we'll resist the prospect of being upstaged and replaced by our own "artificial progeny" (1988: 108).

Moravec's more recent *Robot: Mere Machine to Transcendent Mind* (1999) goes much further in envisioning a spectacular future ecology of Exes, or ex-humans, seeking their destinies in space beyond the limits of our blue earth. But is anything like this plausible?

Caution is needed whenever we speculate about times to come. In their chapter, Stuart Armstrong and Seán ÓhÉigeartaigh investigate the track record of past predictions involving artificial intelligence (AI). Unfortunately, would-be prophets have had little success to date, and there has been scant difference in success rates between experts and non-experts. Armstrong and ÓhÉigeartaigh suggest the use of clearer models, more testable predictions, and a less confident attitude to guessing the future.

That said, change can creep up on us slowly until its full implications start to emerge. Michael Anissimov's chapter considers the likely improvement path of AI and advanced robotics. Anissimov argues that change will be sharp, sudden, and rapid, and will likely produce extraordinary variation of design – especially once powerful artificial intelligences themselves start to participate in AI research programs. In his contribution, neuroscientist/neuroengineer Randal Koene provides an expert and detailed description of current research on the structure and functioning of the brain. Koene believes that uploading of human minds via whole brain emulation should become a reality in the next few decades, and expects to see near-term breakthroughs as researchers develop techniques for large-scale, high-resolution mapping of brain activity.

With our daily experience of "the Cloud" and handy smartphones vastly more capable than the giant "artificial brains" of earlier decades, with war

drones replacing soldiers, and brain scanners that can decode brainwaves and drive prostheses, many of us share a sense of change all around. A sense – perhaps – that *everything* is about to change. We are tantalized by an uncanny and altogether unprecedented prospect: the threat or promise of new kinds of advanced intelligence on our planet.[1]

2 Machines that think?

One of our themes is whether a sufficiently advanced computer could be conscious, and/or be a "self," or (putting it another way) whether some kind of software personality could be "run" on an advanced computer. Increasingly powerful machines that do not possess these characteristics might be very useful tools – they might help us solve problems that have defeated our efforts to date. But if they lack anything resembling consciousness, if they "think" only in the same way as our current computers, much of the charm is lost. To say the least, they will lack the intrigue, allure, and philosophical interest of machines with sensations, emotions, and desires comparable to ours.

There are seemingly strong arguments that our mental states are caused by, or emerge from, complex physical structures and events in our bodies, particularly our brains. Indeed, most philosophers of mind believe that thought and sensation are causally dependent on the structure and functioning of our neural systems.[2] The mystery, however, is just *how* these complex physical systems produce mental states. Consider, for example, your belief that lions are carnivores. Just what is it about the organization of your brain that makes it true of you that you possess this belief (as I'm sure you do)?

If what really matters is that our brains perform a kind of computation, then mental states could be produced by other physical systems devised and programmed to perform the same kind of computation. In that case, the artificial system is not merely simulating but actually *emulating* or *replicating* the relevant functioning of a human brain. By contrast, a computer model of a tornado, however accurate and detailed, does not possess the twister's ability to wreak havoc in the real world. What matters for the destructive power of the tornado is the movement of air particles at high speeds, interacting with trees, motor cars, houses, and whatever else might get in their path. A computer simulation of a tornado does nothing of the kind.

Hence, if computation is what really matters for mental states and mental functioning, then these are organizationally invariant across all sufficiently powerful computational systems. The same inner experiences that you or I have, perhaps of sensation, desire, or puzzlement, could also exist in computers made from very different materials. If computation is all that's

required, we could devise a "functional isomorph" of a human brain – and its functioning would produce the same mental states as the original. On this approach, a stream of mental states is multiply realizable: it could be realized by all functionally isomorphic (though otherwise very different) systems. By contrast, if something else is important – perhaps something to do with the brain's physical composition or the physical products of its activity – we are back with the general truth that simulation is not replication. In that case, what I've called full-scale machine intelligence, complete with sensation and other conscious experience, will not be achieved via computer simulation of a human brain's structures and functioning.

For several decades now, John Searle has argued that the kind of information-processing carried out by digital computers cannot, by itself, cause such things as sensations, beliefs, desires, and thoughts. Without a genuine mind (such as a human one) to assign an interpretation to it, a computer's output is not intrinsically meaningful; it is not *about* anything in particular until we interpret it. Searle's conclusion is that "mental states are biological phenomena." Thus: "Consciousness, intentionality, subjectivity and mental causation are all a part of our biological life history, along with growth, reproduction, the secretion of bile, and digestion" (Searle 1984: 41).

The ongoing debate between proponents of organizational invariance and biological theorists of consciousness is represented in this volume by the respective chapters of two distinguished professors of philosophy: David Chalmers and Massimo Pigliucci. Chalmers characterizes the debate in these terms:

> Biological theorists of consciousness hold that consciousness is essentially biological and that no nonbiological system can be conscious. Functionalist theorists of consciousness hold that what matters to consciousness is not biological makeup but causal structure and causal role, so that a nonbiological system can be conscious as long as it is organized correctly.

But even this description of the debate is controversial. Pigliucci, for one, challenges it. He points out that biological theorists do not strictly insist that no non-biological system can ever be conscious. Nor, strictly speaking, do they deny that what matters is the causal structure and causal role of a system and its elements. They deny something more specific: that a machine can genuinely think and be conscious merely in undertaking the information-processing tasks possible to a digital computer. Despite their emphasis on biology, theorists such as Searle and Pigliucci concede that some unknown kind of physical system might turn out to have similar causal powers to those of the brain.

This leaves a mystery as to just what is crucial, for present purposes, about the brain and its functioning. Moreover, it is not clear why thought or sensation should require particular material (contrast the tornado, which