VERNON PRESS
COGNITIVE SCIENCE AND PSYCHOLOGY

Edited by

Steven S. Gouveia

Contributors include:
Daniel Dennett, Ben Goertzel,
Natasha Vita-More, David Pearce,
Roman V. Yampolskiy
and Stevan Harnad

THE AGE OF ARTIFICIAL INTELLIGENCE

An Exploration

Daniel C. Dennett, Tufts University, USA

THE AGE OF ARTIFICIAL INTELLIGENCE

AN EXPLORATION

Edited by
Steven S. Gouveia
University of Minho, Portugal

Cognitive Science and Psychology
VERNON PRESS

Daniel C. Dennett, Tufts University, USA

In the Americas:
Vernon Press
1000 N West Street,
Suite 1200, Wilmington,
Delaware 19801
United States

In the rest of the world:
Vernon Press
C/Sancti Espiritu 17,
Malaga, 29006
Spain

Cognitive Science and Psychology

Library of Congress Control Number: 2020931461

ISBN: 978-1-62273-872-4

Cover design by Vernon Press using elements designed by FreePik.

Daniel C. Dennett, Tufts University, USA

TABLE OF CONTENTS

Daniel C. Dennett, Tufts University, USA

LIST OF ACRONYMS

ACM = Agent cognitive model

ADM = Agent dialogue model

AGI = Artificial General Intelligence

AI = Artificial Intelligence

AIML = Artificial intelligence markup language

AISE = Artificial intelligence safety engineering

ALAMO = Atelier de Littérature Assistée par la Mathématique et les Ordinateurs

ALICE = Artificial linguistic internet computer entity

APMA = Amino-3-hydroxy-5-methyl-4-isoxazole propionic acid

AKM = Agent knowledge model

A.R.T.A = Atelier de Recherches et Techniques Avancées

BCI = Brain-computer interfaces

BDI = Belief–desire–intention model

BNC = British national corpus

BSC = Bodily self-consciousness

CSs = Constructional schemes

CD = Compact disc

CERN = European organization for nuclear research

CEV = Coherent extrapolated volition

CGs = Conceptual graphs

CL = Computational linguistics

CNS = Central nervous system

COGUI = Conceptual graphs user interface

CP = Cognitive phenomenology

CREB = cAMP response element-binding protein

CRISPR = Clustered regularly interspaced short palindromic repeats

CS = Computational structures

CSA = Conversational software agent

DARPA = Defense advanced research projects agency

DDD = Distributed, decentralized, and democratic

DLT = Distributed ledger technology

DM = Discourse management

DNA = deoxyribonucleic acid

DIY = Do it yourself

DRM = Digital rights management

DRT = Discourse representation theory

ELO = Electronic literature organization

EMI = Experiments in musical intelligence

EQ = Embodiment question

FAI = Friendly artificial intelligence

FEP = Free-energy principle

FOOM = Recursively self-improving artificial intelligence engendered singularity

FOS = Free and open source

GANs = Generative adversarial networks

GM = Grand masters (of chess)

GOFAI = Good old-fashioned artificial intelligence

GPS = Global positioning system

HIS = Hazardous intelligent software

HOT = higher-order thought theory of consciousness

HS = Hazardous software

IBM = International business machines

IEEE = Institute of electrical and electronics engineers

IQ = Intelligence quotient

IoT = Internet of things

KL = Kullback-Leiber

KR = Knowledge representation

LS = Logical structure

LSc = Layered structure of the clause

LSDB = Liveset database

MAI = Malevolent artificial intelligence

MEG = Magnetoelectroencephalography

MIRI = Machine intelligence research institute

MS = Mental states

MT = Machine translation

MW = Musical work

NL = Natural language

NLP = Natural language processing

NLU = Natural-language understanding

NGO = Non-governmental organization

NOC = Non-ordinary consciousness

NCC = Neural correlates of consciousness

NSA = National security agency

OR = Ockham's razor

Orch-OR = Orchestrated objective reduction

PCR = Polymerase chain reaction

PHEN = Phenomenological independence implies metaphysical independence

POS = Part-of-speech

QHOT = Quotational higher-order thought theory of consciousness

R&D = Research and development

RDF = Resource description framework

RNA = Ribonucleic acid

RRG = Role and reference grammar

SA = Simulation argument

SAC = Speech act constructions

SAT = Speech act theory

SH = Simulation hypothesis

SP = Sensory processing

STAGE = Science, technology and governance in europe

TOC = Theatre of consciousness

TOTE = Test-operate-test-exit

TREC = Text retrieval conferences

TT = Turing test

WSD = Word sense disambiguation

XML = Extensible markup language

LIST OF FIGURES

LIST OF TABLES

Daniel C. Dennett, Tufts University, USA

Daniel C. Dennett, Tufts University, USA

INTRODUCTION

"Computers are useless because they can only give you answers"
(Pablo Picasso)[1]

"Chess... is eminently and emphatically the philosopher's game"
(Paul Morphy)[2]

On May 11th, 1997, in the emblematic city of New York, on the 9th floor of the Equitable Center, a major event happened that can be considered as one of the most decisive moments for humanity:[3]

1.e4 c6 2.d4 d5 3.Cc3 dxe4 4.Cxe4 Cd7 5.Cg5 Cgf6 6.Ad3 e6 7.C1f3 h6 8.Cxe6 De7 9.O-O fxe6 10.Ag6+ Rd8 11.Af4 b5 12.a4 Ab7 13.Te1 Cd5 14.Ag3 Rc8 15.axb5 cxb5 16.Dd3 Ac6 17.Af5 exf5 18.Txe7 Axe7 19.c4 1-0.

If you don't know the history of some of the most famous chess games, you will not immediately recognize this sequence. But if you do, you will notice that this is the sum-up of all the 19 moves made in the sixth game of the second match between Garry Kasparov (one of the most brilliant chess players in the world - alongside M. Carlsen, B. Fisher, V. Anand, A. Karpov and P. Morphy) and the computer created by International Business Machines (IBM) named DeepBlue. In the 8th move, after Kasparov played h6 – something he would regret later – DeepBlue decided to sacrifice a knight for a pawn, playing e6, something that Kasparov would never have expected.

What happened next was the pure dominance of the computer, forcing Kasparov to resign on the 19th move and to lose the match.[4] Kasparov would

[1] Different versions of this are cited in William Fifield's original interview with Picasso, "Pablo Picasso: A Composite Interview," published in the Paris Review 32, Summer–Fall 1964and in Fifield's 1982 book *Search of Genius*, New York: William Morrow.

[2] As quoted in *Testimonials to Paul Morphy*, Presented at University Hall, New York, May 25, 1859 (cf. https://play.google.com/books/reader?id=aEZAAAAAYAAJ).

[3] *Time* magazine had claimed something similar regarding the first game of the first match between Kasparov and DeepBlue in 1996. Kasparov had lost the first game (although he would ended up winning the match): the first-game defeat was more than "world historical. It was species-defining".

[4] The match was composed by six gamesand the result after the first five games was a tie, 2½–2½. The match's result was, after the win of DeepBlue in the sixth and last game, 3½–2½.

never have played h6 against a human opponent, nevertheless, since he was playing against a computer, he chose that move: “I didn't want to play. I was sorry about my decision to play h6. Normally computers don't take on e6”.

Although DeepBlue was built with 256 co-processors capable of calculating approximately 200 million positions per second, for Kasparov, the reason for his loss was simple: no computer would use a tactical move such as sacrificing a knight so early in the game. At the time, there were some suspicions that the research team behind the development of DeepBlue was being helped (live in action) by Grand Masters (GM), namely, Bobby Fisher.

In the late eighteenth century, there was a famous chess-playing mechanical automaton called “the Turk[5]”. It was an engraved figure made of wood that could move its pieces and it could play a competent chess game. The “Turk” was branded as the very first artificial system. Touring through America and Europe, it played against professional players, including renowned historical celebrities such as Napoleon Bonaparte and Benjamin Franklin, who were themselves chess aficionados. Of course, this was a hoax, an elaborate, but nevertheless, a fake artificial system with a person cleverly hidden inside the wood structure playing all the moves (cf. Kasparov, 2017: 7).

IBM had one purpose only: to prove that they could build a machine that could defeat the best chess player in the world, the reigning champion. The team responsible for the development of DeepBlue was composed by Murray Campbell (IBM, Thomas J. Watson Research Center), Joel Benjamin (Chess GM and consultant), Feng-hiung Hsu (who started developing DeepBlue while he was at Carnegie Mellon University), Thomas Anantharaman as well as a few others, all managed by Chung-Jen Tan, who was known as the spokesman and the “resident philosopher” of the team.

Kasparov had won the first match against DeepBlue in the previous year, 15 months prior in Philadelphia - this was the second one. The preparation for this second match was difficult because he couldn’t study previous games played by DeepBlue since there were none: he had to play against a black box without any chance of studying and analyzing previous games made by the computer. Worse than that, there was a clause in the contract for the second match that Kasparov completely overlooked: the machine could be rebooted during or after each game. This would make the post-analysis of its games impossible, therefore, eliminating any chances of studying specific games or to recognize any patterns in DeepBlue’s approach to chess.

[5] The machine was nicknamed the Turk because it played its moves through a turbaned marionette attached to a cabinet (cf. Campbell, 1997: 83).

Newsweek's cover called this match "The Brain's Last Stand". The match was covered in all newspapers and broadcasted live on television: the world was seeing firsthand that artificial machines could potentially surpass a game in which human beings had excelled for centuries. After this first conquest, Artificial Intelligence research focused on building computers that could defeat humans in other more complex games, like Go (with AlphaGo), Scrabble (with Quackle) or Jeopardy (with Watson). The artificial system won in every one of these games.

The 1997 match was announced as Kasparov representing Humanity versus the Machine. If he lost the match, then everything could be achievable for AI: it would officially give rise to the "Age of Artificial Intelligence". For contemporary science, chess was seen as the ultimate test of intelligence. In cinema, there are also many chess references along these lines: for example, in Stanley Kubrick's *2001: a Space Odyssey*, there is a scene where the computer HAL9000 plays a chess game against Frank Poole - the game is an actual recreation of a tournament game from 1913 between A. Roesch and W. Schlage (cf. Campbell, 1997: 79).

It's very interesting to notice that the birth of Computer Science and AI is associated with the first reflections and thoughts about creating a machine that could play chess. One of the first discussions happens in Charles Babbage's *The Life of a Philosopher* (1845). In 1945, one century later, Konrad Zuse describes a program that could generate legal moves in chess. Claude Shannon, founder of information theory, wrote a paper titled "A Chess-Playing Machine" (1950) where he describes two main approaches in building a competent chess-program: Type A was about creating a program based on brute-force, that is, a program that could analyze and calculate in seconds all the millions of potential positions for a specific move; and Type B, a program that would have a strategic and goal-focused approach, more like humans' beings play chess. Shannon believed that the Type B approaches would be more successful in beating a human, but it was eventually the Type A approach which won the race, culminating in the development of DeepBlue.

Alan Turing also thought about whether an artificial system that could play chess, called "Turochamp", famously known as the "Turing's paper machine". The program, created in 1952,[6] was written with his colleague David Champernowne and was composed by a specific set of instructions. The only problem is that there were no digital computers in that time, so Turing wrote

[6] See the original publication here: see https://en.chessbase.com/post/reconstructing-turing-s-paper-machine).

the program by hand - it was later recreated in an actual computer, where it turned out to not be a very competent chess player.

In 1950, one of the founders of Artificial Intelligence, Herbet Simon, stated it would only take 10 years for a machine to become the world champion of chess – he was just mistaken by 30 years (cf. Cambpell, 1997: 83-85).

In 1956, John McCarthy, owning to Alexander Kronrod, described chess as the "drosophila of Artificial Intelligence" (cf. McCarthy, 1990). Like the common fruit fly, which juxtaposed to more complex organic systems, its research is quite "simple" to do. Nevertheless, it can produce significant knowledge about further complex systems. Just as at its base, chess is a "simple" game, and therefore, it can teach us an ample amount about human cognition and intelligence (cf. Ensmenger, 2012).

As we can see, there was a great, almost obsessive focus on chess in the first days of AI research because it was believed that chess was the ultimate pinnacle of human intelligence. Nowadays, this seems, a tiny bit exaggerated. At present, Artificial Intelligence is focused on what is called "weak AI": it excels at very specific tasks – like playing games, facial recognition or driverless cars – but it is not even close to achieving human-level intelligence. The reason is quite simple: Artificial Intelligence's research methods are more about imitating human performances – the Turing Test is a very good example of this idea – than to look for its own achievements and goals.

Consequently, all the tasks we can describe and codify can be outperformed by machines. But the real achievement of a fully conscious machine seems far still. Because we do not know anything about consciousness, Artificial Intelligence conceived as "strong AI", that is, a conscious A.I., may never be fully achieved. For that, we need the right theoretical framework – we need better and more philosophical research.

The book is divided into five main sections. Section I is dedicated to reflections on the Intelligence of AI and will open with a chapter by Soenke Ziesche and Roman V. Yampolskiy, which discuss the mathematics of intelligence for grouped minds, nested minds as well as deducted minds. The following chapter, by Stevan Harnad, debates if the existence of feelings is a real caveat for a system that would pass the Turing's Test. Next, Daniel Dennett argues against the mysterianism position that we cannot study our conscious mind and explain why AI, although theoretically possible to be achieved, may never be practically accomplished because of its costs and the lack of epistemic advantages of such an achievement. Finally, closing Section I, David Pearce discusses three different ways to connect human and artificial intelligence: by fusion, replacement or co-evolution, arguing that only the third process may be plausible.

Section II follows, dedicated to discussion on the relationship between consciousness, simulation and artificial intelligence. Gabriel Axel Montes and Ben Goertzel present the concept of a 'mindplex' as a way of enhancing the connection between human and artificial minds; for which, they use the concept of non-ordinary consciousness (NOC) and show how that perspective can be relevant for understanding the mind and cognition in general. Cody Turner follows offering two arguments in favor of the thesis that a phenomenology of cognition is neither reducible to, nor dependent upon, sensory phenomenology. If this thesis is plausible, then it follows that AI consciousness may not require embodiment to be emulated, as commonly assumed. The next chapter by Nicole Hall argues that aesthetic experience is a fundamental feature of human consciousness and separates human from artificial intelligence. She argues further that it is a mistake to confuse the mere possibility of achieving "conscious singularity", as she defines it, with human consciousness and its capacity for aesthetically experiencing natural environments. To conclude Section II, Steven S. Gouveia introduces an intriguing idea that we may be living in a computer simulation, briefly debating the main reasons in favor and against this hypothesis.

Section III, dedicated to aesthetical creativity and language in artificial intelligence, opens with a chapter by Caterina Moruzzi where in the light of recent developments in AI music software generators discusses the question, "Can a computer create a musical work?" On the same topic, René Mogensen proposes a formal representation of content in computational creativity of music, noting that in order to achieve complete computational music creativity, aesthetic experience appears to be necessary. Mariana Chinellato Ferreira follows, applying the same discussion about aesthetical creativity in computer-generated literature, analyzing specific software such as PoeTryMe. Closing Section III, Kulvinder Panesar presents a functional linguistic perspective on natural language processing in artificial intelligence.

The subsequent Section IV is on the Ethics of the Bionic Brain Peter A. DePergola II opens by offering the argument that neurocognitive enhancement can be justified as morally plausible if it (a) promotes general moral character, (b) complements human nature and (c) effects a deeper sense of individual and social identity. Next, Tomislav Miletić and Frederic Gilbert discuss the potential harms of brain-computer interfaces (BCI) on self-determination, warning that any patient who accepts the use of such future AI medical technology should be sufficiently prepared for the symbiotic relation before the implementation. Following on the same topic of the ethics of BCI, Aníbal M. Astobiza, Txetxu Ausin, Ricardo M. Ferrer and Stephen Rainey focus on some issues raised by BCI research, identifying some dangers, challenges and opportunities for the elaboration of a common ethical and legal framework concerning issues of

safety, ethics and data protection. To conclude Section IV, Natasha Vita-More argues that Artificial General Intelligence (AGI) can be used as a tool to improve our knowledge about ourselves and the world.

Finally, to close the book, Section V follows on the Ethics of Artificial Intelligence, starting with a chapter by Federico Pistono and Roman V. Yampolskiy that provides some general guidelines for the creation of a Malevolent Artificial Intelligence (MAI) with the goal of challenging the AI Safety Community to continue its effort by discovering and reporting specific problems. Following, Hasse Hämäläinen attempts to find the most plausible answer to the question of whether a machine could be attributed moral and legal rights and obligations, arguing that if a machine can perform a specific task or set of capacities as human beings do, then the rights and obligations of humans should also be applied to machines. The next chapter by Vernon Vinge discusses the ethical implications of the Singularity, offering a set of ethical guidelines to avoid the extinction of human race. Finally, to conclude both Section V and the book, Eray Özkural discusses the ethical and epistemological implications of the Free Energy Principle: the idea that a self-organization occurs through minimization of free energy.

The Age of Artificial Intelligence is imminent, if it's not already here. We should make sure that we invest in the right people and the right ideas in order to create the best solutions possible. My hope is that this book will help to influence the right minds. If Reason killed god in the 20th and 21st century, Reason – philosophy, science and technology – may resurrect it in form of an Artificial General Intelligence: an AI that may know everything about anything. We should make sure that we create the right kind of god and that we keep it in the right hands.

I would like to finish this introduction by deeply thanking all the people who made this project feasible.[7]

Steven S. Gouveia
Ottawa, 10/09/2019

[7] I would like to acknowledge the precious help offered by Jessica Clarke and, to Susan Schneider for her valuable feedback.

References

Campbell, Murray S. (1997) "'An Enjoyable Game': How HAL Plays Chess" In *HAL's Legacy: 2001's Computer as Dream and Realtiy* (David G. Stock, ed.), Cambridge, MA: MIT Press.

Ensmenger, Nathan (2012) "Is chess the drosophila of artificial intelligence? A social history of an algorithm" *Soc Stud Sci.*, 42 (1): 5-30.

Feng-hsiung, Hsu, Anantharaman, Thomas, Campbell, Murray and Nowatzyk, Andreas (1990) "A Grandmaster-level Chess Machine" *Scientific American*, 263 (4): 44-51.

Fifield, William (1982) *Search of Genius*, New York: William Morrow

Frey, Peter W. (ed.) (1983) *Chess Skill in Man and Machine*, New York: Springer-Verlag.

Hsu, Feng-hsiung (2002) *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*, Princeton, NJ: Princeton University Press.

Kasparov, Gary (2017) *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, New York: PublicAffairs™.

Levy, David and Newborn, Monty (1991) *How Computers Play Chess*, New York, NY: Computer Science Press.

Marsland, Anthony T. and Schaeffer, Jonathan (eds.) (1990) *Computers, Chess and Cognition*, New York, NY: Springer-Verlag.

McCarthy, John (1990) "Chess as the Drosophila of A.I.". In *Computers, Chess and Cognition* (Marsland T.A., Schaeffer J., eds.), New York, NY: Springer, pp. 227- 237.

Shannon, Claude (1950) "Programming a computer for playing chess" *Philosophical Magazine*, 41: 256-75.

CHAPTER 3

THE AGE OF POST-INTELLIGENT DESIGN[1]

Daniel C. Dennett

Tufts University, USA

1. What are the limits of our comprehension?

> "If the brain were so simple we could understand it, we would be so simple we couldn't."
>
> - Emerson M. Pugh, *The Biological Origin of Human Values.*[2]

Human comprehension has been steadily growing since prehistoric times. For forty millennia and more, we have been living in the age of intelligent design – crafting pots, tools, weapons, clothes, dwellings and vehicles; composing music and poetry; creating art; inventing and refining agricultural practices; and organizing armies, with a mixture of dutiful obedience to tradition, heedless and opportunistic improvisation, and knowing, intentional, systematic R&D, irregularly punctuated with moments of "inspired" genius. We applaud intelligent design in all arenas, and aspire from infancy to achieve recognition for our creations. Among the artifacts we have created is the concept of God, the Intelligent Designer, in our own image. That's how much we value the intelligent designers in our societies.

We recognize the value of these fruits of our labors, and our laws and traditions have been designed to create an artificial environment in which we can preserve and enhance our accumulated wealth. It is a perfectly real environment, not a merely virtual world, but it is no less an artifact, and we call it civilization. We are well aware that our species is no more immune to extinction than any other, and that we might all expire in a plague or

---

[1] "The Age of Post-Intelligent Design", from FROM BACTERIA TO BACH AND BACK: THE EVOLUTION OF MINDS by Daniel C. Dennett. Copyright © 2017 by Daniel C. Dennett. Used by permission of W. W. Norton & Company, Inc.

[2] Quoted by George E. Pugh, *The Biological Origin of Human Values* (1978), p. 154. In a footnote, the author writes that this quote was from his father, Emerson Pugh, around 1938. The quote is also widely attributed to Lyall Watson, born in 1939. http://todayinsci.com/P/Pugh_Emerson /PughEmerson-Quotations.htm.

technological catastrophe, or – only slightly less dire – we might destroy civilization and revert to the "state of nature," as Hobbes called it, where life is nasty, brutish, and short. But has it ever occurred to us that this age of comprehending heroes might simply expire while Homo sapiens went right on reproducing, outliving its name, the knowing hominin? There are some unsettling signs that we are becoming overcivilized, thanks to the ingenuity of all the labor-saving inventions on which we have become so dependent, and are entering the age of post-intelligent design.

The epigraph to this chapter, Pugh's clever reflection on the audacious project of using our brains to understand our brains, has been attributed, in different versions, to many authors, and it may well have been independently reinvented many times. A variant species includes one of my favorite George Carlin one-liners:

> For years and years and years, I thought my brain was the most important organ of my body, until one day I thought, hmm. Look who's telling me that!

Is there an important truth lurking here, or is this just another way Cartesian gravity has of diverting us from the quest to understand human consciousness? Noam Chomsky (1975) has proposed a distinction that has drawn a lot of attention and converted a few disciples: on the one hand there are *problems*, which we can solve, and on the other hand there are *mysteries*, which we can't. Science and technology have solved many problems about matter, energy, gravity, electricity, photosynthesis, DNA, and the causes of tides, tuberculosis, inflation, and climate change, for instance. Progress is being made on thousands of other *problems*. But no matter how advanced our scientific problem-solving becomes, there are problems that are beyond human comprehension altogether, which we might better call mysteries. Consciousness tops Chomsky's list of mysteries, along with free will. Some thinkers – now known as *mysterians* – have been eager to take this unargued claim on his authority and run with it. There may possibly be mysteries systematically beyond the ken of humanity now and forever, but the argument in favor of this disheartening conclusion put forward by Chomsky and the other mysterians, while superficially appealing, is not persuasive. Here is a rendering of the Argument from Cognitive Closure, drawing on various versions:

> It is an undeniable fact of biology that our brains are strictly limited, like the brains of all other creatures. From our relatively Olympian vantage point, we can see that fish are clever in their way but obviously not equipped to understand plate tectonics, while dogs draw a blank when it comes to the concept of democracy. Every brain must suffer from cognitive closure (McGinn, 1991) with regard to a host of issues

> that are simply beyond it, unimaginable and unfathomable. We don't have a miraculous res cogitans between our ears, but just lots of brain tissue subject to the laws of physics and biology.

So far, so good. I have no objection to anything in this beginning, which articulates uncontroversial facts about the physical world. But then it goes on:

> It would be profoundly unbiological – wouldn't it? – to suppose that our human brains were somehow exempt from these natural limits. Such delusions of grandeur are obsolete relics from our prescientific past.

This would be compelling if it weren't for the equally obvious biological fact that human brains have become equipped with addons, thinking tools by the thousands, that multiply our brains' cognitive powers by many orders of magnitude. Language, as we have seen, is the key invention, and it expands our individual cognitive powers by providing a medium for uniting them with all the cognitive powers of every clever human being who has ever thought. The smartest chimpanzee never gets to compare notes with other chimpanzees in her group, let alone the millions of chimpanzees who have gone before. The key weakness of the Argument from Cognitive Closure is the systematic elusiveness of good examples of mysteries. As soon as you frame a question that you claim we will never be able to answer, you set in motion the very process that might well prove you wrong: you raise a topic of investigation. While your question may get off on the wrong foot, this fact about it is likely to be uncovered by the process of trying to answer it. The reflexive curiosity of philosophy – going "meta" about every question asked – is almost a guarantee that there will be something approximating exhaustive search – sometimes no better than random, sometimes brilliantly directed – of variations on the question that might prove more perspicuous. Asking better and better questions is the key to refining our search for solutions to our "mysteries," and this refinement is utterly beyond the powers of any languageless creature. "What is democracy?" A dog will never know the answer, to be sure, but it will never even understand the question. We can understand the questions, which radically changes our quest, turning unimaginable mysteries into problems worth trying to solve.

Perhaps in consideration of this practically limitless power of language to extend our grasp, Chomsky has recently moderated his position (2014). While there is a "conceptual distinction" between problems and mysteries, "we accept the best explanations science can give us" even when we can't imagine how they work. "It doesn't matter what we can conceive any more. We've given up on that." In other words, thanks to language, and the tools of science it makes possible, we can have a good scientific theory of some perplexing phenomenon, a theory worth endorsing, while not *really* understanding it.

That is, we could be justified in accepting it, even betting our lives on the truth of its predictions, while not understanding how or why it works. Whether or not this revision would appeal to the mysterians, it is still an interesting idea. But could it be true?

Downloading thousands of culturally acquired thinking tools may permit us to magnify our powers dramatically, but doesn't that just postpone cognitive closure? How much schooling can an individual mind/brain absorb? Here we notice an ambiguity in the mysterian surmise. Is it the claim that there are mysteries that *no single human* mind can comprehend or that there are mysteries that are beyond the *pooled* comprehension of whole civilizations? The idea of distributed comprehension – the idea that *we as a group* might understand something that none of us individually could fully understand – strikes some people as preposterous, so loyal are they to the ideal of the intelligent do-it-yourself designer, the genius who has it all figured out. This is a *motif* with many familiar variations. A painting by the *studio* of Rembrandt is less valuable, less a masterpiece, than a painting by Rembrandt himself. Novels almost always have solo novelists (the hardworking editors who reshape the penultimate draft often do not even get recognized), and when creative teams succeed – Gilbert and Sullivan, Rodgers and Hammerstein – they almost always involve a division of labor: one does the lyrics and one does the music, for instance. But coauthored works of nonfiction have been common for centuries, and, in the sciences today, there are fields in which a single-authored paper is a rarity.

One of the founding documents of cognitive science, *Plans and the Structure of Behavior* (1960) was written by George Miller, Eugene Galanter, and Karl Pribram. Its introduction of the idea of a TOTE unit (Test-Operate-Test-Exit) was an early formalization of the idea of feedback loops, and it played an important role in the transition from behaviorism to cognitive modeling. For all its early influence, it is seldom read these days, and a joke once commonly heard was that Miller wrote it, Galanter formalized it, and Pribram believed it. The very idea that such a division of labor might be possible – and successful – was risible at the time, but no longer. Science is full of collaborations in which theoreticians – who understand the math – and experimentalists and fieldworkers – who rely on the theoreticians without mastering the math – work together to create multiple-author works in which many of the details are only partially understood by each author. Other combinations of specialized understanding flourish as well.

So let's imagine a multi-author, multi-volume3 book, *The Scientific Theory of Consciousness*, that comes to be uncontroversially accepted by the scientific community. The volumes become, if you like, the standard textbooks on human consciousness, used in courses across neuroscience, psychology, philosophy, and other fields where consciousness is an important phenomenon – but although some intrepid souls claim to have read through the whole boxed set, nobody claims to have mastered all the levels of explanation. Would it count as vindicating Chomsky's mysterianism – consciousness is still a mystery, since no single theorist can really conceive of it – or as knocking over yet another of the mysterians' candidates for unfathomable mysteries?

We have learned, as civilization has progressed, that a division of labor makes many things possible. A single person, or family, can make a simple house or canoe, and a small community can raise a barn or a stockade, but it takes hundreds of workers with dozens of different talents to build a cathedral or a clipper ship. Today peer-reviewed papers with hundreds of coauthors issue from CERN and other bastions of Big Science. Often none of the team members can claim to have more than a bird's-eye-view comprehension of the whole endeavor, and we have reached a point where even the most brilliant solo thinkers are often clearly dependent on their colleagues for expert feedback and confirmation.

Consider Andrew Wiles, the brilliant Princeton mathematician who in 1995 proved Fermat's Last Theorem, a towering achievement in the history of mathematics. A close look at the process he went through, including the false starts and unnoticed gaps in the first version of his proof, demonstrates that this triumph was actually the work of many minds, a community of communicating experts, both collaborating and competing with each other for the glory, and without the many layers of achieved and battle-tested mathematics on which Wiles's proof depended, it would have been impossible for Wiles or anyone else to judge that the theorem had, in fact, been proven.[4] If you are a lone wolf mathematician and think you have proved Fermat's Last Theorem, you have to consider the disjunction: *Either I have just proved Fermat's Last Theorem or I am going mad*, and since history shows that many

[3] In "Belief in Belief," chapter 8 in Breaking the Spell (2006), I discuss the division of labor between scientists in different fields who rely on each other's expertise ("they do the understanding and we do the believing!") and point out that in theology, the point is often made that nobody understands the terms of the discussion.

[4] Simon Singh, "The Whole Story," http://simonsingh.net/books/fermats -last-theorem/the-whole-story/, is by far the most accessible account I have found, an edited version of an essay published in *Prometheus* magazine.

brilliant mathematicians have been deluded in thinking they had succeeded, you have to take the second alternative seriously. Only the formal concession and subsequent congratulations of your colleagues could or should put that anxiety to rest.

Even artists, poets, and musicians, treasured for their individuality and "divine afflatus," do best when they have an intimate working knowledge and understanding of the works of their predecessors. The twentieth-century rebels who made something of a fetish of defying "the canon," attempting to create works of hyper-originality, are either fading into oblivion or proving that the staying power of their creations is due to more appreciation of their traditions than they were willing to admit. The painter Philip Guston once eloquently acknowledged his indirect dependence on all he had extracted and digested from the intelligent design of others:

> I believe it was John Cage who once told me, "When you start working, everybody is in your studio – the past, your friends, enemies, the art world, and above all, your own ideas – all are there. But as you continue painting, they start leaving, one by one, and you are left completely alone. Then, if you're lucky, even you leave" (Guston, 2011: 30).

What kind of limits are there on the brains we were born with? For now, we can note that whether the limits are practical or absolute, we have found, and largely perfected, a work-around that postpones the confrontation with our frailty: collaboration, both systematic and informal. Groups can do things, and (arguably) understand things, that individuals cannot, and much of our power derives from that discovery. It is possible to resist this idea of group comprehension, but only – so far as I can see – by elevating comprehension to a mystical pinnacle that has little or nothing to do with the comprehension we rely on, in ourselves and others, to solve our problems and create our masterpieces. This blunts the edge of the mysterian argument. By ignoring the power of collaborative understanding, it raises an obsolete issue, viewing comprehension as an all-or-nothing blessing, which it seldom, if ever, is.

Descartes, in his day, was very concerned to secure *perfect* comprehension for his "clear and distinct" ideas, and for this, he argued, he needed to prove the existence of a benign, all-powerful, non-deceiving God. His thought-experimental hypothesis was that there might otherwise be an evil demon bent on deceiving him about his most confidently held convictions, and this "possibility in principle" fixed his method – and tied his hands quite securely. For Descartes, only the kind of certainty we reserve for dead-obvious mathematical truths (2 + 2 = 4, a plane triangle has three straight sides and interior angles adding up to two right angles) was good enough to count as *real* knowledge, and only the crystalline comprehension we can have of the individual steps of a

maximally simple proof could count as *perfect* comprehension. Where Descartes relied on God as the guarantor of his proofs, today we rely on the improbability of multiple thinkers arriving, by different routes, at the *same* wrong result. (Note that this is an application of the principle that dictated taking at least three chronometers on your sailing ship, so that when they began to disagree on what time it was, the odd one out was very probably wrong.) We tend to overlook the importance of the fact that we have voluminous experience of many people independently coming up with the same answer to multiplication and division questions, for instance, but if that were not our experience, no amount of analytic reflection on the intrinsic necessity of mathematics – or the existence of a benign God – would convince us to trust our calculations. Is arithmetic a sound system of calculation? *Probably* – so very probably that you can cheerfully bet your life on it.

2. "Look Ma, no hands!"

> "Civilization advances by extending the number of important operations we can perform without thinking about them."
>
> —Alfred North Whitehead

> "What I cannot create, I do not understand."
>
> —Richard Feynman

I have argued that the basic, bottom-up, clueless R&D done by natural selection has gradually created cranes – labor-saving products that make design work more effective – which have opened up Design Space for further cranes, in an accelerating zoom into the age of intelligent design, where top-down, reflective, reason-formulating, systematic, foresighted R&D can flourish. This process has succeeded in changing the balance of selective forces that shape us and all other organisms and in creating highly predictive theories that retrospectively explain the very processes of their own creation. This cascade of cranes is not a miracle, not a gift from God, but a natural product of the fundamental evolutionary process, along with the other fruits of the Tree of Life.

To review, over several thousand years, we human beings have come to appreciate the powers of individual minds. Building on the instinctive habits of all living things, we distinguish food from poison, and, like other locomoting organisms, we are extra sensitive to animacy (guided movement) in other moving things, and more particularly to the beliefs and desires (information and goals) that guide those movements, tracking as best we can *who knows what* and *who wants what*, in order to guide our own efforts at hide and seek. This native bias is the genetic basis for the intentional stance, our practice of treating each other as rational agents guided by largely true beliefs and largely well-ordered desires. Our uninterrupted interest in these

issues has generated the *folk psychology* that we rely on to make sense of one another. We use it to predict and explain not just the repetitive behaviors we observe in our neighbors and ourselves, and the "forced moves" that anyone would be stupid not to execute, but even many of the strokes of "insight" that are the mark of "genius." That is, our expectations are very frequently confirmed, which cements our allegiance to the intentional stance, and when our expectations are confounded, we tend to fall back on "explanations" of our failure that are at best inspired guesswork and at worst misleading mythmaking.

We encourage our children to be curious and creative, and we self-consciously identify the ruts and boundaries in our own thinking processes so that we can try to overcome them. The minds we prize most are the minds that are neither too predictable (boring, unchallenging) nor too chaotic. *Practice makes perfect*, and we have invented games that encourage us to rehearse our mind-moves, such as chess and Go and poker, as well as prosthetic devices – telescopes, maps, calculators, clocks, and thousands of others – that permit us to apply our mind-moves in ever more artificial and sophisticated environments. In every sphere of inquiry and design, we have highly organized populations of experts collaborating to create and perfect theories and other artifacts, and we have adopted traditions and market mechanisms to provide the time, energy, and materials for these projects. We are the intelligent designers living in a world intelligently designed for intelligent designers by our ancestors. And now, after centuries of dreaming about this prospect, we have begun designing and producing artifacts that can design and produce artifacts (that can design and produce artifacts …).

Many hands make light work. That's another adage that is as applicable to mind-work as to barn-raising, but we are now discovering that hands-off design work is often not only lighter and easier, but, thanks to the artifacts we have recently designed, more – in a word – competent. Nanotechnology, the new and burgeoning field of chemistry and materials science that is beginning to construct artifacts atom by atom, has featured the brilliant and patient handiwork of pioneers who have developed sophisticated tools for manipulating (moving, cutting, isolating, immobilizing, etc.) bits of matter at the nanometer scale (a nanometer is one-billionth of a meter). Like GOFAI before it, nanotechnology began as top-down intelligent design, a brilliant method for *hand making* large inventories of "miracle drugs, "smart materials," and other nanorobots. It has had triumphs and will surely have many more, especially with the new nanotool of CRISPR at its disposal (see, for a brief nontechnical introduction, Specter, 2015). Like PCR (polymerase chain reaction), the technique that revolutionized gene sequencing, CRISPR, which permits genes to be edited and spliced together more or less ad lib, replaces highly sophisticated

and laborious techniques, a labor-saving invention that reduces the time and effort required by orders of magnitude. Jennifer Doudna of UC Berkeley and Emmanuelle Charpentier, now of the Max Planck Institute, are two of the supremely intelligent designers of this new crane.

These techniques, like those developed by Pixar and other computer-animation companies, create push-button automated processes that replace thousands of days of *brilliant drudgery* (not an oxymoron – extremely talented people doing extremely repetitive but demanding work). When Walt Disney Productions released *Snow White and the Seven Dwarfs* in 1937, it astonished the world with its lifelike animations, the fruits of the labors of hundreds of talented animators, working in highly organized teams to solve the problems of getting lifelike action, complete with all the jiggle and bounce of reality, onto thousands of cells, or frames, of film. Those heroic artistic sweatshops are historical relics now; the talents one needed to be that kind of frame-by-frame animator are now largely obsolete, and the same is true about the talents of early molecular biologists who ingeniously isolated gene fragments and patiently coaxed them to divulge their sequences, one codon at a time. Similar tales of the automation of heretofore tedious intellectual tasks could be told about other fields, from astronomy to textual analysis. In general, these tasks amount to gathering, sorting, and refining data on a large scale, giving the human data-interpreter more free time to reflect on the results. (I will never forget the time I spent a day in the laboratory of a promising young neuroscientist gathering data from experiments on macaques with chronically implanted electrodes in their brains. Late in the day I asked him a question about his views on a theoretical controversy then boiling about the role of activity in various brain areas on modulating consciousness; he sighed and replied, "I don't have time to think! I'm too busy running experiments.") The new techniques that minimize the brilliant drudgery are amazingly competent, but they are still tools – not robotic colleagues – utterly dependent on the decisions and intentions of intelligent tool users and directors – lab directors and studio directors.

Today, however, we are beginning to appreciate, and exploit, the truth of Orgel's Second Rule: Evolution is cleverer than you are. The bottom-up, tireless algorithms of natural selection (and their close cousins) are being harnessed by intelligent designers in many fields to do the dirty work of massive searches, finding precious needles in all kinds of haystacks. Some of this exploration involves actual biological natural selection in the laboratory. For instance, Frances Arnold, at Caltech, does award-winning protein engineering, creating novel proteins by breeding them, in effect. She has devised systems for generating huge populations of variant genes – DNA

recipes for proteins – and then testing the resulting proteins for talents never before found in Nature.

> We are developing new tools for protein engineering and using them to create new and improved catalysts for carbon fixation, sugar release from renewable polymers such as cellulose, and biosynthesis of fuels and chemicals (Arnold, 2013).

What she recognized was that since the space of *possible* proteins is Vastly greater than the space of *existing* proteins, there are almost certainly traversable paths of gradual evolution that have never yet been explored to destinations that would provide us with wonder drugs, wonder tissues, and wonder catalysts, a host of nanorobots that can do our bidding once we find them. When she was a graduate student, a senior scientist warned her that there were no known proteins that had anything like the properties she was hoping to obtain. "That's because there's never been selection for them" was her intrepid reply.

> Consequently, these enzymes may open up whole new regions of "chemical space" that could not be explored in previous medicinal chemistry efforts (Arnold, 2013).

Frances Arnold has created a technology for generating novel proteins – long sequences of amino acids that, when linked together, fold into evolved nanorobots[5] with remarkable powers. A strikingly different technology developed by David Cope[6], emeritus professor of music at University of California at Santa Cruz, uses a computer program to generate novel music – long sequences of notes and chords that, when linked together, yield musical compositions with remarkable powers: imitation Bach, imitation Brahms, Wagner, Scott Joplin, and even musical comedy songs (cf. Cope and Hofstadter, 2001). How "original" are the thousands of compositions churned out by Cope's EMI (Experiments in Musical Intelligence)? Well, they are clearly derivative and involve heavy borrowing from the great composers whose styles they mimic, but they are nonetheless not mere copies, and not mere

[5] Evolving robots on the macroscale have also achieved some impressive results in very simplified domains, and I have often spoken about the work in evolutionary robotics by Inman Harvey and Phil Husbands at Sussex (e.g., Harvey et al. 1997), but I have not discussed it in print.

[6] Cope's Virtual Music (2001) includes essays by Douglas Hofstadter, composers, musicologists, and me: "Collision Detection, Muselot, and Scribble: Some Reflections on Creativity." The essays are filled with arresting observations and examples, and the volume includes many musical scores and comes with a CD.

copies with a few random mutations; they are much better than that. They involve taking in and digesting the works of the masters and extracting from that computational process the core, the gist, the style of that composer, and then composing novel pieces in that style, a very sophisticated musical feat. (Try it, if you are a musician, and see: compose a piano piece that is pure Chopin or Mozart – or Count Basie or Erroll Garner. Simple parody or caricature is not that hard, especially of a jazz pianist as mannered as Erroll Garner, but composing good music requires significant musical insight and talent – in a human composer.)

Experiments in Musical Intelligence, designed and improved by Cope over more than three decades, has produced many well- constructed piano pieces, songs, symphonies, and other compositions, all untouched by Cope's editorial hand, except for the final aesthetic judgment of which of the bounty most deserve to be heard. I arranged for a nice test of EMI – one of many that have been conducted over the years – at the Montreal Bach Festival in December of 2015, where I gave a talk summarizing some of the main points of this book, and closing with a performance, by Ukrainian pianist Serhiy Salov, of four short piano pieces. I told the audience of over 300 Bach lovers that at least one was by Bach and at least one was by EMI, and after they were played, the audience voted (with eyes closed). Two EMI inventions were declared genuine Bach by dozens in the audience – maybe not a majority in either case, but close – and when I asked those who had got them all right to stand, only about a dozen rose to a round of applause.

Cope, like Arnold, sets the goals and decides when to declare victory but otherwise is hands off. These very different research projects are thus variations on Darwin's theme of *methodical selection*, in which the selective force of natural selection is focused through the nervous system of a discerning, purposeful, foresighted agent. But the heavy lifting is left to the inexorable pattern-finding powers of the algorithms of natural selection, in cascades of uncomprehending generate-and-test cycles that gradually refine the search process.

Since natural selection is a substrate-neutral[7] family of algorithms that can occur in any medium with a few simple properties, evolution in silico (simulated in a computer program) is sometimes faster and cheaper than evolution *in vivo*, and can be applied to almost any question or problem you formulate. Pedro Domingos's recent *book The Master Algorithm* (2015) is a lively and authoritative survey of all the new varieties of Darwinian and – shall we say – *Darwinesque*

[7] See DDI (1995) on substrate neutrality.

systems of "machine learning" or "deep learning." Domingos simplifies the stampede by identifying five "tribes of machine learning": symbolists (the descendants of GOFAI); connectionists (the descendants of McCulloch and Pitts's logical neurons – see Dennett, 2017: 110); evolutionaries (John Holland's genetic algorithms and their offspring); Bayesians (those who have devised practical algorithms for achieving the competences of hierarchical networks of Bayesian expectation-generators); and analogizers[8] (the descendants of the nearest-neighbor algorithm invented by Fix and Hodges [1951]). In different ways, all five echo the pattern of natural selection. Obviously, being computer-based, they all are ultimately composed of Turing's simplest comprehension- free competences (conditional branching and arithmetic), and except perhaps for the creations of the symbolists, they are bottom-up, needle-in-haystack-finding repetitive churnings that gradually, with great reliability, home in on good (or good enough) answers to daunting problems.

John Holland, the beloved and recently deceased mentor of dozens of brilliant cognitive scientists and computer scientists at the Santa Fe Institute and the University of Michigan, invented genetic algorithms, where the parallels with evolution by natural selection are obvious (and delicious to Darwinians): there is the generation of huge populations of variant codings, which are each given the opportunity to make progress on solving a problem, with the winners of this environmental test getting to reproduce (complete with a sort of sex, and "crossover" like the random gene-mixing that occurs in the creation of our sperm and ova). Over many generations, the competence of the initially randomly concocted strings of computer code is multiplied and refined. Genetic algorithms have been used to design the fascinating evolved virtual creatures of Karl Sims (see the many websites devoted to this serious playground of imagination) and also such no-nonsense engineering triumphs as circuit boards and computer programs. Domingos notes (Domingos, 2015: 133) that in 2005, a patent was issued for a genetically designed factory-optimization system (General Leslie Groves, they are closing in on you). Architects[9] have begun using genetic algorithms to optimize the functional

[8] See also Douglas Hofstadter's many works on the importance of analogy finding, especially *Metamagical Themas* (1985), *Fluid Concepts and Creative Analogies* (1995), and *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking* (2013), coauthored with Emmanuel Sander (first published in French as *L'Analogie. Cœur* de la pensée; published in English in the United States in April).

[9] For some examples of the use of genetic algorithms in architecture, see Sullivan-Fedock (2011), Asadia et al. (2014), Yu et al. (2014), and for optimizing electoral redistricting, see Chou et al., (2012).

properties of buildings – for instance, their strength, safety, use of materials, and use of light and energy.

In scientific research, machine learning is being harnessed to solve, by brute force, problems that are simply beyond human analysis. It is noteworthy that the late Richard Feynman, brilliant theoretical physicist that he was, spent many of his last days exploring the use of supercomputers to solve problems in physics that defied his wizardry with equations. And he lived to see his maxim rendered more or less obsolete. While it may still be true that what you cannot create you cannot understand, creating something is no longer the guarantee of understanding that it used to be. It is now possible to make – very indirectly – things that do what we want them to do but which we really cannot understand. This is sometimes called black-box science. You buy the latest high-tech black box, feed in your raw data, and out comes the analysis; the graphs are ready to print and publish, yet you couldn't explain in detail how it works, repair it if it broke, and *it is not clear that anybody else could either*. This possibility was always staring us in the face, of course; things we "make by hand" (boats, bridges, engines, symphonies) we can (more or less) control as we construct, understanding each step along the way. Things we "make the old-fashioned way" (children, grandchildren,) defy our comprehension because of our obliviousness to the details of the processes that create them. Today, we are generating brain-children, and brain-grandchildren, and brain-greatgrandchildren that depend on processes we cannot follow in detail, even when we can prove that the results are trustworthy.

The use of computers in research has generated several quite distinct kinds of challenges to Feynman's maxim. Some mathematical proofs executed (in part or entirely) by computer are simply too long for a single human mathematician to check each step, which has been, for a good reason, the standard of acceptance for several thousand years. What should give? A famous case is the computer-assisted proof in 1976 of the four-color theorem first discussed by Möbius in the 1840s: Any map of contiguous areas separated by shared boundaries can be colored in with just four colors such that the same color never appears on both sides of a boundary. After many failed proofs by some of the world's cleverest mathematicians, Kenneth Appel and Wolfgang Haken harnessed a computer to deal with and dismiss the nearly 2,000 different possibilities that had to be ruled out, as they themselves had proven. For some years, their proof was not generally accepted because it seemed to involve a humanly uncheckable series of steps by the computer, but the wide consensus among mathematicians today is that this is a proven theorem. (And alternative proofs have since been constructed, also using computers.) This was an "intuitive" result: nobody had managed to produce a clear counterexample in spite of person-centuries of trying, so most mathematicians figured it was true long before it was proven.

But there are also counterintuitive theorems that have been proven with the help of computers. In chess, for instance, the fifty-move rule, which declared any game a draw that proceeded for fifty moves without a capture or pawn move, was long viewed by experts as more than generous, but it was upset by the discovery – thanks to computer analysis – of some mating nets (winning series of moves by one side that cannot be escaped once entered), that involve no captures or pawn moves and exceed fifty moves by hundreds of moves. After some experimentation with revising the number, it was officially decided by FIDE, the international governing body of the game, to keep the fifty-move rule, since it was a possibility in principle that would never come up in serious (human) play.

The computer programs that analyze chess positions, like those that prove mathematical propositions, are traditional, top-down, intelligently designed programs. The programs Domingos is mainly concerned with are strikingly different. As he puts it, "We can think of machine learning as the inverse of programming, in the same way that the square root is the inverse of the square, or integration is the inverse of differentiation" (Domingos, 2015: 7). Yet another strange inversion of reasoning, or better, another instance of the basic Darwinian inversion: competence without comprehension. The "central hypothesis" of Domingos's book is beyond audacious:

> All knowledge—past, present, and future—can be derived from data by a single, universal learning algorithm. I call this learner the Master Algorithm. If such an algorithm is possible, inventing it would be one of the greatest scientific achievements of all time. In fact, the Master Algorithm is the last thing we'll ever have to invent because, once we let it loose, it will go on to invent everything else that can be invented. All we need to do is provide it with enough of the right kind of data, and it will discover the corresponding knowledge (Domingos, 2015: 25).

It isn't clear if he really means it, since he soon backpedals:

> OK, some say, machine learning can find statistical regularities in data, but it will never discover anything deep, like Newton's laws. It arguably hasn't yet, but I bet it will (Domingos, 2015: 39).

A wager, then, not a hypothesis he thinks he can secure by reasoned argument at this time in his book. In any case, it's useful to have his articulation of this extreme prospect, since no doubt many people have half-formed nightmares about just such eventualities, and it will help shine some skeptical light on them. We can begin with the claim Domingos responds to with his wager. Can machine learning ever advance beyond finding "statistical regularities"? Domingos bets that it will, but what is the basis for his optimism?

3. The structure of an intelligent agent

We have seen how Bayesian networks are excellent at teasing out the statistical regularities that matter to the organism – its affordances. Animal brains, equipped by natural selection with such networks, can guide the bodies they inhabit with impressive adroitness, but by themselves have scant ability to *adopt novel perspectives*. That, I have argued, requires an infestation of memes, cognitive competences (habits, ways) designed elsewhere and installed in brains, habits that profoundly change the cognitive architecture of those brains, turning them into minds, in effect. So far, the only animals whose brains are thus equipped are *Homo sapiens*.

Just as the eukaryotic cell came into existence in a relatively sudden instance of *technology transfer*, in which two independent legacies of R&D were united in a single stroke of symbiosis to create a big leap forward, the human mind, the *comprehending* mind, is – and had to be – a product of symbiosis, uniting the fruits of two largely independent legacies of R&D. We start, I have argued, with animal brains that have been, to a considerable extent, redesigned to be excellent bases for thinking tools designed elsewhere – memes. And chief among them, words. We *acquire* most of our words unconsciously, in this sense: we were not aware of learning seven new words a day when we were young, and for most words – words that aren't explicitly introduced to us – we only gradually home in on their meanings thanks to unconscious processes that find the patterns in our early experience of these words. Once we *have* the words, we can begin *using* them, but without necessarily noticing what we are doing. (For every word in your vocabulary, there was a debutante token, the first time you used it either in a public speech act or an internal monologue or musing. How often have you been aware of doing that with the new words that have entered your working vocabulary in, say, the last decade? Ever?) Once words become our familiar tools, not mere sounds associated with contexts, we can start using them to create new perspectives on everything we encounter.

So far, there are few signs of this sort of phenomenon emerging in the swiftly growing competences of deep-learning machines. As Domingos stresses, learning machines are (very intelligently) designed to avail themselves of Darwinesque, bottom-up processes of self-redesign. For IBM's Watson, the program that beat champion contestants Ken Jennings and Brad Rutter in the *Jeopardy* television quiz program in 2011, the words it was competent to string together into winning answers were not *thinking tools* but just nodes located in a multidimensional space of other nodes, not so much memes as fossil traces of human memes, preserving stupendous amounts of information about human beliefs and practices without themselves being active participants in those practices. Not yet, but maybe someday. In short, Watson

doesn't yet think thoughts using the words about which it has so much statistical information. Watson can answer questions (actually, thanks to *Jeopardy*'s odd convention, Watson can compose questions to which the *Jeopardy* clues are the answers: *Jeopardy*: "The capital of Illinois," contestant: "What is Springfield?"), but this is not having a conversation.

It is the capacity to self-monitor[10], to subject the brain's patterns of reaction to yet another round (or two or three or seven rounds) of pattern discernment, that gives minds their breakthrough powers.[11] In the current environment of machine learning, it is the human users, like Frances Arnold in her protein workshop and David Cope with his Experiments in Musical Intelligence, the designers and other operators of the machines, who are in position to play those roles, evaluating, adjusting, criticizing, tweaking, and discarding the dubious results that often emerge. They are the critics whose quality control activities provide the further selective forces that could "in principle" raise these systems into comprehension, promoting them from tools into colleagues, but that's a giant step or series of giant steps. From this perspective, we can see more clearly that our meme-infested minds harbor users, critics of the raw deliverances of our animal brains without which we would be as naïve as other mammals, who are wily on their own turf but clueless in the face of serious novelty.

Curiosity killed the cat, according to one meme, and animal curiosity, driven bottom-up by the presence of novelty, is an important high-risk, high-payoff feature in many species, but only human beings have the capacity for controlled, systematic, foresighted, hypothesis-testing curiosity, a feature of the users that emerge in each brain, users who can exploit their brains' vast capacity for uncovering statistical regularities. The user-illusion of consciousness plays the same role in each of us that the human-computer interfaces of Watson and other deep-learning systems play; they provide something like a showcase for talents, a "marketplace of ideas" in which real-time evaluation and competition can enhance the speed and resolution of quality control.

---

[10] I discuss this design strategy in an unpublished paper, "A Route to Intelligence: Oversimplify and Self-monitor" (1984) that can be found on my website: http://ase.tufts.edu/cogstud/dennett/recent.html.

[11] Watson, notably, does have some specialized self-monitoring layers: it must assess its "confidence" in each candidate answer and can also adjust its threshold of confidence, taking more or less risk in answering. I don't want to sell Watson short; it is a multilayered, multitalented system.

So human *thinking* – as Darwin recognized in the phenomenon he called *methodical selection* – can speed up natural selection by focusing selective power through the perceptual and *motivational* systems of domesticators. Frances Arnold isn't just *farming* her proteins; she is doing intensive, directed *breeding* of the new proteins. That should alert us to the prospect that our marvelous minds are not immune to fads and fancies that bias our self-redesign efforts in bizarre and even self-defeating ways. Like the preposterous plumage encouraged into existence by methodical pigeon fanciers, and the pathetic disabilities patiently engineered into various "toy" dog varieties, human beings can – often with the help of eager accomplices – shape their minds into grotesque artifacts that render them helpless or worse.

This suggests – but certainly does not prove – that without us machine *users* to interpret the results, critically and insightfully, deep-learning machines may grow in competence, surpassing animal brains (including ours) by orders of magnitude in the bottom-up task of finding statistical regularities, but never achieve (our kind of) comprehension. "So what?" some might respond. "The computer kind of bottom-up comprehension will eventually submerge the humankind, overpowering it with the sheer size and speed of its learning." The latest breakthrough in AI, AlphaGo, the deep-learning program that has recently beaten Lee Seedol, regarded by many as the best human player of Go in the world, supports this expectation in one regard if not in others. I noted that Frances Arnold and David Cope each play a key quality-control role in the generation processes they preside over, as critics whose scientific or aesthetic judgments decide which avenues to pursue further. They are, you might say, piloting the exploration machines they have designed through Design Space. But AlphaGo itself does something similar, according to published reports: its way of improving its play is to play thousands of Go games against itself, making minor exploratory mutations in them all, *evaluating* which are (probably) progress, and using those evaluations to adjust the further rounds of practice games. It is just another level of generate and test, in a game that could hardly be more abstract and insulated from real-world noise and its attendant concerns, but AlphaGo is learning to make "intuitive" judgments about situations that have few of the hard-edged landmarks that computer programs excel at sorting through. With the self-driving car almost ready for mass adoption – a wildly optimistic prospect that not many took seriously only a few years ago – will the self-driving scientific exploration vehicle be far behind?

So practical, scientific, and aesthetic judgment may soon be offloaded or outsourced to artificial agents. If Susan Blackmore is right, this abdication or alienation of human judgment is already being pioneered in the digital world of popular music and Internet memes – *tremes*, in her new terminology (see

Dennett, 2017: 237). There has been a surfeit of memes for centuries, with complaints dating to the earliest days of the printing press, and ever since then, people have been willing to pay for filters, to strain out the time-wasting, mind-clogging, irritating memes one way or another. Do not try to read every poem by every poet; wait until some authoritative poet or critic puts forth a highly selective anthology. But which authority should you trust? Which meets your needs and tastes? You can subscribe to a literary journal that regularly reviews such volumes, along with the work of individual poets. But which literary journal should you trust? Check out their reputations as reported in still other journals you can buy. There is a living to be made catering to the expressed needs of individual meme seekers, and if business slacks off, you can try to create a new need which you can then service. All very familiar. But we are entering a new era where the filters and second-guessers and would-be trendsetters may not be people at all, but artificial agents. This will not suit everybody, as we will see in the next section. But that may not stop hierarchical layers of such differential replicators from burgeoning, and then we may indeed face the calamity encountered by the Sorcerer's Apprentice and the multiplying brooms.

In an IBM television advertisement, Watson, "in conversation" with Bob Dylan, says that it can "read 800 million pages a second." Google Translate, another icon among learning machines, has swept aside the GOFAI systems that were top-down attempts to "parse" and interpret (and thereby understand, in at least a pale version of human comprehension) human language; Google Translate is an astonishingly swift, good – though still far from perfect – translator between languages, but it is entirely parasitic on the corpus of translation that has already been done by human bilinguals (and by volunteer bilingual informants who are invited to assist on the website). Prospecting for patterns, sifting through millions of passages that have already been well translated (well enough to be found online), Google Translate settles into a likely (probable) acceptable translation *without any actual comprehension at all.*

This is a contentious claim that requires some careful unpacking. There is a joke about an Englishman who says, "The French call it a *couteau*, the Italians call it a *cotello,* the Germans call it a *Messer*, and we English call it a *knife* – which, after all, is *what it is!*" The selfsatisfied insularity of the Englishman is attached to something he knows – what a knife is – that has no counterpart (it seems) in the "knowledge" of Google Translate. In the jargon of cognitive science, the Englishman's knowledge of the meaning of "knife" (and "*couteau*" and the other synonymous terms) is *grounded* in nonlinguistic knowledge, acquaintance, familiarity with knives, with cutting and sharpening, the heft and feel of a carving knife, the utility of a pen knife, and so forth. The Englishman has, with respect to

the word *knife*, what you probably do not have with respect to the English word *snath*, even if you know that the Germans call it a *Sensenwurf*. But hang on. Google Translate no doubt has a rich body of data about the contexts in which "knife" appears, a neighborhood that features "cut," "sharp," "weapon" but also "wield," "hold," "thrust," "stab," "carve," "whittle," "drop," and "bread," "butter," "meat" and "pocket," "sharpen," "edge," and many more terms, with their own neighborhoods. Doesn't all this refined and digested information about linguistic contexts amount to a sort of grounding of the word "knife" after all? Isn't it, in fact, the only sort of grounding most of us have for technical terms such as "messenger RNA" and "Higgs boson"? It does guide the translation process down ever more appropriate channels. If you rely on Google Translate to be your bilingual interpreter, it will hardly ever let you down. Doesn't that demonstrate a serious degree of comprehension? Many will say NO! But if we are to keep this adamant denial from being a mere ritualistic defiance of the machine, there had better be something the *real* comprehender can *do* with his or her (or its) comprehension that is beyond the powers of Google Translate.

Maybe this will do the trick: It is one thing to *translate* a term paper from English to French, and another thing to *grade* that term paper. That won't do, because Thomas Landauer's pioneer development of "latent semantic analysis" (see, e.g., Littman et al. 1998) has already created a computer program that does precisely that (Rehder et al. 1998). A professor sets an essay question on an exam, and writes an A+ answer to the question, which is then given to both the computer program and a human teaching assistant as an example of what a good essay on the topic should be. (The A+ answer is *not* shown to the examination takers, of course.) Then the program and the teaching assistant grade all the student answers, and the program's grades are in closer agreement to the professor's judgments than the grades submitted by the teaching assistant, who is presumably a budding expert in the field. This is unnerving, to say the least; here is a computer program that doesn't understand English, let alone the subject matter of the course, but simply (!) on the basis of sophisticated statistical properties exhibited by the professor's model essay evaluates student answers to the same questions with high reliability. Assessment competence without comprehension! (Landauer has acknowledged that *in principle* a student could contrive an essay that was total nonsense but that had all the right statistical properties, but any student who could do that would deserve an A+ in any case!)

Then how about the task of simply *having a sensible conversation with a human being?* This is the classic Turing Test[12], and it really can separate the

[12] For an analysis and defense of the Turing Test as a test of genuine comprehension, see

wheat from the chaff, the sheep from the goats, quite definitively. Watson may beat Ken Jennings and Brad Rutter, two human champions in the TV game *Jeopardy*, but that isn't a free-range conversation, and the advertisements in which Watson chats with Jennings or Dylan or a young cancer survivor (played by an actress) are scripted, not extemporaneous. A real, open-ended conversation between two speaking agents is, as Descartes (1637) observed in his remarkably prescient imagining of a speaking automaton, a spectacular exhibition of great – if not *infinite*, as Descartes ventured to say – cognitive skills. Why? Because ordinary human conversation is conducted in a space of possibilities governed by Gricean free-floating rationales! I may not expressly *intend* that you *recognize* my intention to get you to believe that what I am saying is true (or that it is irony, or kidding, or transparent exaggeration), but if you are not up to that kind of recognition, and if you are also not up to originating speech acts having similar free-floating rationales that explain your own responses and challenges, you will not be a convincing, or engaging, conversationalist. Grice's iterated layers of cognition may not accurately represent real-time features underlying a *performance*, but they do specify a *competence*.

A participant in a high-powered conversation has to be able to recognize patterns in its own verbal actions and reactions, to formulate hypothetical scenarios, to "get" jokes, call bluffs, change the subject when it gets tedious, explain its earlier speech acts when queried, and so forth. All this requires – if it is not magic – the representation of all the discriminations that must be somehow *noticed* in order to provide the settings for the mental and ultimately verbal actions taken. For instance, if you do not or cannot notice (in some minimal, possibly subliminal sense) that I'm joking, you can't go along with the gag, except by accident. Such noticing is not simply a matter of your animal brain making a discrimination; it is rather some kind of heightened influence that not only retrospectively distinguishes what is noticed from its competitors at the time but also, just as importantly, contributes to the creation of a *noticer*, a relatively long-lasting "executive," not a place in the brain but a sort of political coalition that can be in control over the subsequent competitions for some period of time. Such differences in the aftermath ("And then what happens?") can be striking.

Imagine being asked to complete partial words (the "word stem completion paradigm") and being confronted with

---

my "Can Machines Think?" (1985), reprinted in Brainchildren, with two postscripts (1985) and (1997); "Fast Thinking" in *The Intentional Stance* (1987); and especially "The Chinese Room," in *Intuitions Pump* (2013), where I discuss examples of the cognitive layering that must go into some exchanges in a conversation (Dennett, 2013: 326–327).

sta ___

or

fri ___

What occurred to you? Did you think *start* or *stable* or *station*, for instance, and *frisk*, *fried*, *friend*, or *frigid*? Suppose that, a few seconds before you get the word stem to complete, an answer word is flashed very briefly on the screen, thus:

staple

followed a second later by *sta*___. The temptation to answer "staple" would be huge, of course. But suppose the experimenters said at the outset of the experiment: "If you've just seen a word flash, *don't* use it as the answer!" Then, not surprisingly, you can overcome the urge and say something different most of the time, maybe *stake* or *starlight*. You are unlikely to say "staple" because you can follow the exclusion policy recommended by the experimenter. But that's only if you notice (or are conscious of) the flashed word. If the word is flashed for only 50msec and followed by a "mask" – a patterned screen – for 500msec, you are *more* likely to say "staple" in spite of trying to follow the instruction (Debner and Jacoby, 1994).[13] Notice how clean the design of this experiment is: two groups of subjects, one group told to *use* the "priming" word if it's a good answer, and the other group told *not* to use the "priming" word if it's a good answer. Both groups get primes that are either 50msec or 500msec long, followed by a mask. The mask doesn't mask the 500msec-long primes – subjects notice them, can report them, are conscious of them, and either use them or refrain from using them, as requested. But the mask does mask the 50mseclong primes – subjects claim not to have seen any priming word at all (this is a standard "backward masking" phenomenon). In both kinds of cases, short and long, there is discrimination by the brain of the prime, as shown by the fact that in the exclusion condition, the short-duration primes *raise* the probability of being the answer given, while the long-duration primes *lower* that probability. Dehaene and Naccache (2001) note "the impossibility for subjects [i.e., executives] to strategically use the unconscious information."

My claim, then, is that deep learning (so far) *discriminates* but doesn't *notice*. That is, the flood of data that a system takes in does not have relevance

[13] For more on such experiments, see my "Are We Explaining Consciousness Yet?" (2001) and also Dehaene and Naccache (2001), Smith and Merikle (1999), discussed in Merikle et al. (2001).

for the system except as more "food" to "digest." Being bedridden, not having to fend for itself, it has no goals beyond increasing its store of well-indexed information. Beyond the capacity we share with Watson and other deep learning machines to acquire know-how that depends on statistical regularities that we extract from experience, there is the capacity to *decide* what to search for and *why*, given one's current aims. It is the absence of *practical* reason, of intelligence harnessed to pursue diverse and shifting and self-generated ends, that (currently) distinguishes the truly impressive Watson from ordinary sane people. If and when Watson ever reaches the level of sophistication where it can enter fully into the human practice of reason-giving and reason-evaluating, it will cease to be merely a tool and become a colleague. And then Watson, not just Watson's creators and maintainers, would be eligible for being considered *responsible* for its actions.

The way in which deep-learning machines are dependent on human understanding deserves further scrutiny. In chapter 8, (Dennett, 2017:157–160), we considered Deacon's bold criticism of traditional AI: would-be mind designers who abstract away from the requirements of energy capture and self-protection thereby restrict their search to parasitic systems, always dependent on human maintenance – they are tools, not colleagues. Now we can see that the kind of comprehension AI systems are currently exhibiting – and it is becoming breathtakingly competitive with the best human comprehension – is also parasitic, strictly dependent on the huge legacy of human comprehension that it can tap. Google Translate would be nowhere without the millions of good translations by bilinguals that it draws upon, and Watson's inhumanly compendious factual knowledge is likewise dependent on all those millions of pages it sucks off the Internet every day. To adapt once again Newton's famous remark, these programs stand on the shoulders of giants, helping themselves to all the cleverness on display in the earlier products of intelligent design.

This is nicely illustrated by a problem I set for my students when Watson beat Jennings and Rutter on *Jeopardy*. I gave them the assignment of coming up with questions that they thought would stump Watson but be easy for Jennings or Rutter (or any normal human being). (It's notable that on *Jeopardy*, the rules had to be adjusted in Watson's favor. For instance, the problems set to Watson were all verbal, with no vision or hearing required.) The likely stumpers (in my opinion) involve imagination in one way or another:

> The happy word you could spell on the ground using a cane, a hula hoop, and a slingshot.
> Ans. What is joy?

> Make a small animal huge by changing one letter in its name.
> Ans. What is mouse to moose?
>
> The numeral, between 0 and 9, that would make a good shape for a hot tub and adjacentswimming pool.
> Ans. What is 8?

I have better examples, but I wouldn't publish them – or put them on the Internet – since then Watson would probably sweep them up and keep them on hand for a future contest! Watson doesn't need an imagination when it can poach on the imaginations of others. Note that in this regard Watson is deeply Darwinian: neither Watson nor natural selection depend on foresight or imagination because they are driven by processes that relentlessly and without comprehension extract information – statistical patterns that can guide design improvements – from what has already happened. They are both blind to types of events that haven't happened in the scope of their selection processes. Of course, if there really is nothing new under the sun, this is no limitation, but human imagination, the capacity we have to envision realities that are not accessible to us by simple hill climbing from where we currently are, does seem to be a major game-changer, permitting us to *create*, by *foresighted design*, opportunities and, ultimately, enterprises and artifacts that could not otherwise arise. A conscious human mind is not a miracle, not a violation of the principles of natural selection, but a novel extension of them, a new crane that adjusts evolutionary biologist Stuart Kauffman's concept of the *adjacent possible*: many more places in Design Space are adjacent to us because we have evolved the ability to think about them and either seek them or shun them. The unanswered question for Domingos and other exponents of deep learning is whether learning a sufficiently detailed and dynamic *theory* of agents with imagination[14] and reason-giving capabilities would enable a system (a computer program, a Master Algorithm) to generate and exploit the abilities of such agents, that is to say, to generate all the morally relevant powers of a person.[15]

---

[14] I discuss the prospects of such a powerful theory or model of an intelligent agent, and point out a key ambiguity in the original Turing Test, in an interview with Jimmy So about the implications of Her, in "Can Robots Fall in Love" (2013), *The Daily Beast*. http://www.thedailybeast.com/articles/2013/12/31/can-robots-fall-in-love -and-why-would-they.html.

[15] Spike Jonze's science fiction film, Her (2013), starring Joaquin Phoenix and the voice of Scarlett Johansson as the Siri-like virtual person on his cell phone with whom he falls in love, is one of the two best speculative explorations of this question to date, along with Alex Garland's *Ex Machina* (2015).

My view is (still) that deep learning will not give us – in the next fifty years – anything like the "superhuman intelligence" that has attracted so much alarmed attention recently (Bostrom, 2014; earlier invocations are Moravec, 1988; Kurzweil, 2005; and Chalmers, 2010; see also the annual Edge world question, 2015; and Katchadourian, 2015). The accelerating growth of competence in AI, advancing under the banner of deep learning, has surprised even many professionals in the field, not just long-time commentators and critics such as myself. There is a long tradition of hype in AI, going back to the earliest days, and many of us have a well-developed habit of discounting the latest "revolutionary breakthrough" by, say, 70% or more, but when such high-tech mavens as Elon Musk and such world-class scientists as Sir Martin Rees and Stephen Hawking start ringing alarm bells about how AI could soon lead to a cataclysmic dissolution of human civilization in one way or another, it is time to rein in one's habits and reexamine one's suspicions. Having done so, my verdict is unchanged but more tentative than it used to be. I have always affirmed that "strong AI" is "possible in principle" – but I viewed it as a negligible practical possibility,16 because it would cost too much and not give us anything we really needed. Domingos and others have shown me that there may be feasible pathways (technically and economically feasible) that I had underestimated, but I still think the task is orders of magnitude larger and more difficult than the cheerleaders have claimed, for the reasons presented in this chapter, and in chapter 8 (the example of Newyorkabot, Dennett, 2017: 164).

So I am not worried about humanity creating a race of superintelligent agents destined to enslave us, but that does not mean I am not worried. I see other, less dramatic, but much more likely, scenarios in the immediate future that are cause for concern and call for immediate action.

---

16 When explaining why I thought strong AI was possible in principle but practically impossible, I have often compared it to the task of making a robotic bird that weighed no more than a robin, could catch insects on the fly, and land on a twig. No cosmic mystery, I averred, in such a bird, but the engineering required to bring it to reality would cost more than a dozen Manhattan Projects, and to what end? We can learn all we need to know about the principles of flight, and even bird flight, by making simpler models on which to test our theories, at a tiny fraction of the cost. People have recently confronted me with news items about the latest miniature drones, smaller than robins, as small as flying insects, and asked me if I wanted to reconsider my analogy. No, because they are not autonomous self-controlled robots but electronic marionettes, and besides, they can't catch flies or land on twigs. Maybe someday they will be able to do these things, but only if DARPA wastes billions of dollars. I have discussed the Singularity in "The Mystery of David Chalmers" (2012) and "The Singularity – An Urban Legend?" (2015).

4. What will happen to *us*?

Artifacts already exist – and many more are under development – with competences so far superior to any human competence that they will usurp our authority as experts, an authority that has been unquestioned since the dawn of the age of intelligent design. And when we cede hegemony to these artifacts, it will be for very good reasons, both practical and moral. Already it would be *criminally negligent* for me to embark with passengers on a transatlantic sailboat cruise without equipping the boat with several GPS systems. Celestial navigation by sextant, compass, chronometer, and Nautical Almanac is as quaint a vestige of obsolete competence as sharpening a scythe or driving a team of oxen. Those who delight in such skills can indulge in them, using the Internet to find one another, and we celestial navigators can prudently bring our old-fashioned gear along, and practice with it, on the off chance that we will need a backup system. But we have no right to jeopardize lives by shunning the available high-tech gadgets.

We all still learn the multiplication table up to 12x12 and how to use it for larger numbers (don't we?), and we can do long division problems with pencil and paper, but few know how to execute an algorithm for extracting a square root. So what? Do not waste your effort and brain cells on tasks you can order by pressing a few keys or just asking Google or Siri. The standard response to the worriers is that when educating our children, we do need to teach them the principles of all the methods we ourselves are still adept at using, and for this comprehension, a certain minimal level of actual experience with the methods is practically valuable, but we can (probably) get the principles to sink in without subjecting our children to old-fashioned drudgery. This seems to make good sense, but how far does it generalize?

Consider medical education. Watson is just one of many computer-based systems that are beginning to outperform the best diagnosticians and specialists on their own turf. Would *you* be willing to indulge your favorite doctor in her desire to be an old-fashioned "intuitive" reader of symptoms instead of relying on a computer-based system that had been proven to be a hundred times more reliable at finding rare, low-visibility diagnoses than any specialist? Your health insurance advisor will oblige you to submit to the tests, and conscientious doctors will see that they must squelch their yearnings to be diagnostic heroes and submit to the greater authority of the machines whose buttons they push. What does this imply about how to train doctors? Will we be encouraged to jettison huge chunks of traditional medical education – anatomy, physiology, biochemistry – along with the ability to do long division and read a map? *Use it or lose it* is the rule of thumb cited at this point, and it has many positive instances. Can your children read road maps as easily as you do or have they become dependent on GPS to guide them?

How concerned should we be that we are dumbing ourselves down by our growing reliance on intelligent machines?

So far, there is a fairly sharp boundary between machines that enhance our "peripheral" intellectual powers (of perception, algorithmic calculation, and memory) and machines that at least purport to replace our "central" intellectual powers of comprehension (including imagination), planning, and decision-making. Hand calculators; GPS systems; Pixar's computer graphics systems for interpolating frames, calculating shadows, adjusting textures and so forth; and PCR and CRISPR in genetics are all quite clearly on the peripheral side of the boundary, even though they accomplish tasks that required substantial expertise not so long ago. We can expect that boundary to shrink, routinizing more and more cognitive tasks, which will be fine *so long as we know where the boundary currently is*. The real danger, I think, is not that machines more intelligent than we are will usurp our role as captains of our destinies, but that we will over-estimate the comprehension of our latest thinking tools, prematurely ceding authority17 to them far beyond their competence.

There are ways we can reinforce the boundary, even as we allow it to shrink, by making it salient to everyone. There are bound to be innovations that encroach on this line, and if recent history is our guide, we should expect each new advance to be oversold. There are antidotes, which we should go to some lengths to provide. We know that people are quick to adopt the intentional stance toward anything that impresses them as at all clever, and since the default assumption of the intentional stance is rationality (or comprehension), positive steps should be taken to *show* people how to temper their credulity when interacting with an anthropomorphic system. First, we should expose and ridicule all gratuitous anthropomorphism in systems, the cute, ever-more-human voices, the perky (but canned) asides. *When you are interacting with a computer, you should know you are interacting with a computer*. Systems that deliberately conceal their shortcuts and gaps of incompetence should be deemed fraudulent, and their creators should go to jail for committing the crime of creating or using an artificial intelligence that impersonates a human being.

We should encourage the development of a tradition of hyper-modesty, with all advertising duly accompanied by an obligatory list of all known limits, shortcomings, untested gaps, and other sources of cognitive illusion (the way we now oblige pharmaceutical companies to recite comically long lists of

[17] I enlarge upon this concern in "Information, Technology, and the Virtues of Ignorance" (Daedalus, 1986), reprinted in Brainchildren (1998) and in "The Singularity – An Urban Legend?" (2015).

known side effects whenever they advertise a new drug on television). Contests to expose the limits of comprehension, along the lines of the Turing Test, might be a good innovation, encouraging people to take pride in their ability to suss out the fraudulence in a machine the same way they take pride in recognizing a con artist. Who can find the quickest, surest way of exposing the limits of this intelligent tool? (Curiously, the tolerance and politeness we encourage our children to adopt when dealing with strangers now has the unwanted effect of making them gullible users of the crowds of verbalizing non-agents they encounter. They must learn that they should be aggressive and impolitely inquisitive in dealing with newly encountered "assistants.")

We should hope that new cognitive prostheses will continue to be designed to be parasitic, to be tools, not collaborators. Their only "innate" goal, set up by their creators, should be to respond, constructively and transparently, to the demands of the user. A cause for concern is that as learning machines become more competent at figuring out what we, their users, probably intend, they may be designed to conceal their "helpful" extrapolations from us. We already know the frustration of unwanted automatic "correction" of what are deemed typographical errors by spell-checkers, and many of us disable these features since their capacity to misconstrue our intentions is still too high for most purposes. That is just the first layer of semi-comprehension we have to deal with. There are already stress lines developing around current developments that call for comment. Google has a program for enhancing their search engine by automatically figuring out what it suspects the user *really* meant by entering the input symbol string.[18] This would no doubt be useful for many purposes, but not for all. As Douglas Hofstadter has noted, in an open letter to a former student, then at Google working on this project:

> It worries me and in fact deeply upsets me that Google is trying to undermine things that I depend on on a daily basis, all the time.
> When I put something in quotes in a Google search, I always mean it to be taken literally, and for good reason. For example (just one type of example among many), as a careful writer, I am constantly trying to figure out the best way of saying something in one language or another, and so I will very frequently check two possible phrasings against each other, in order to see whether one has a high frequency and the other a very low frequency. This is an extremely important way for me to find things out about phrasings. If Google, however, doesn't take my phrasings literally but feels free to substitute other words willy-nilly

[18] Cf. http://googleblog.blogspot.com/2010/01/helping-computers-understandlanguage.html

> inside what I wrote, then I am being royally misled if I get a high count for a certain phrase. This is very upsetting to me. I want machines to be reliably mechanical, not to be constantly slipping away from what I ask them to do. Supposed "intelligence" in machines may at times be useful, but it may also be extremely unuseful and in fact harmful, and in my experience, the artificial intelligence (here I use the word "artificial" in the sense of "fake," "non-genuine") that these days is put into one technological device after another is virtually always a huge turnoff to me.
>
> I am thus not delighted by what your group is doing, but in fact greatly troubled by it. It is just one more attempt to make mechanical devices not reliable as such. You ask Google to do X, presuming that it will do precisely X, but in fact it does Y instead, where Y is what it "thinks" you meant. To me, this kind of attempt to read my mind is fantastically annoying if not dangerous, because it almost never is correct or even in the right ballpark. I want machines to remain reliably mechanical, so that I know for sure what I am dealing with. I don't want them to try to "outsmart" me, because all they will do in the end is mislead and confuse me. This is a very elementary point, and yet it seems to be being totally ignored at Google (or at least in your group). I think it is a very big mistake (Personal correspondence, 2010, with Abhijit Mahabal).

At the very least, such systems should (1) prominently announce when they are trying to be "mind readers" not merely "mechanical" and (2) offer users the option of turning off the unwanted "comprehension" in the same way you can turn off your all-too-enterprising spell-checker. A "strict liability" law might provide a much-needed design incentive: anyone who *uses* an AI system to make decisions that impact people's lives and welfare, like users of other dangerous and powerful equipment, must be trained (and bonded, perhaps) and held to higher standards of accountability, so that it is always in their interests to be extra-scrupulously skeptical and probing in their interactions, lest they be taken in by their own devices. This would then indirectly encourage designers of such systems to make them particularly transparent and modest, since users would shun systems that could lead them down the primrose path to malpractice suits.

There is another policy that can help keep the abdication of our cognitive responsibilities in check. Consider technology for "making us stronger": on the one hand, there is the bulldozer route, and on the other hand, the Nautilus machine route. The first lets you do prodigious feats while still being a 98-pound weakling; the second makes you strong enough to do great things on your own. Most of the software that has enhanced our cognitive powers has been of the bulldozer variety, from telescopes and microscopes to genome-sequencers and

the new products of deep learning. Could there also be Nautilus-type software for bulking up the comprehension powers of individuals? Indeed there could be, and back in 1985, George Smith and I, along with programmers Steve Barney and Steve Cohen, founded the Curricular Software Studio at Tufts with the aim of creating "imagination prostheses," software that would furnish and discipline students' minds, opening up notorious pedagogical bottlenecks, allowing students to develop fluent, dynamic, accurate models in their imagination of complex phenomena, such as population genetics, stratigraphy (interpreting the geological history of layers of rock), statistics, and the computer itself. The goal was to make systems that, once mastered, could be set aside, because the users had internalized the principles and achieved the level of comprehension that comes from intensive exploration. Perhaps it is now time for much larger projects designed to help people think creatively and accurately about the many complex phenomena confronting us, so that they can be independently intelligent, comprehending users of the epistemological prostheses under development, not just passive and uncritical beneficiaries of whatever technological gifts they are given.

We have now looked at a few of the innovations that have led us to relinquish the *mastery of creation* that has long been a hallmark of understanding in our species. More are waiting in the wings. We have been motivated for several millennia by the idea expressed in Feynman's dictum, "What I cannot create, I do not understand." But recently our ingenuity has created a slippery slope: we find ourselves indirectly making things that we only partially understand, and they in turn may create things we do not understand at all. Since some of these things have wonderful powers, we may begin to doubt the value – or at least the preeminent value – of understanding. "Comprehension is so *passé*, so *vieux jeux*, so old-fashioned! Who needs understanding when we can all be the beneficiaries of artifacts that save us that arduous effort?"

Is there a good reply to this? We need something more than tradition if we want to defend the idea that comprehension is either intrinsically good – a good in itself, independently of all the benefits it indirectly provides – or practically necessary if we are to continue living the kinds of lives that matter to us. Philosophers, like me, can be expected to recoil in dismay from such a future. As Socrates famously said, "the unexamined life is not worth living," and ever since Socrates we have taken it as self-evident that achieving an ever-greater understanding of *everything* is our highest professional goal, if not our highest goal absolutely. But as another philosopher, the late Kurt Baier, once added, "the over-examined life is nothing to write home about either." Most people are content to be the beneficiaries of technology and

medicine, scientific fact-finding and artistic creation without much of a clue about how all this "magic" has been created. Would it be so terrible to embrace the *over*-civilized life and trust our artifacts to be good stewards of our well-being?

I myself have been unable to concoct a persuasive argument for the alluring conclusion that comprehension is "intrinsically" valuable – though I find comprehension to be one of life's greatest thrills – but I think a good case can be made for preserving and enhancing human comprehension *and* for protecting it from the artifactual varieties of comprehension now under development in deep learning, for deeply *practical* reasons. Artifacts can break, and if few people understand them well enough either to repair them or substitute other ways of accomplishing their tasks, we could find ourselves and all we hold dear in dire straits. Many have noted that for some of our high-tech artifacts, the supply of repair persons is dwindling or nonexistent. A new combination color printer and scanner costs less than repairing your broken one. Discard it and start fresh. Operating systems for personal computers follow a similar version of the same policy: when your software breaks or gets corrupted, do not bother trying to diagnose and fix the error, un-mutating the mutation that has crept in somehow; reboot, and fresh new versions of your favorite programs will be pulled up from safe storage in memory to replace the copies that have become defective. But how far can this process go?

Consider a typical case of uncomprehending reliance on technology. A smoothly running automobile is one of life's delights; it enables you to get where you need to get, on time, with great reliability, and for the most part, you get there in style, with music playing, air conditioning keeping you comfortable, and GPS guiding your path. We tend to take cars for granted in the developed world, treating them as one of life's constants, a resource that is always available. We plan our life's projects with the assumption that of course a car will be part of our environment. But when your car breaks down[19], your life is seriously disrupted. Unless you are a serious car buff with technical training, you must acknowledge your dependence on a web of tow-truck operators, mechanics, car dealers, and more. At some point, you decide to trade in your increasingly unreliable car and start afresh with a brand-new model. Life goes on, with hardly a ripple.

But what about the huge system that makes this all possible: the highways, the oil refineries, the automakers, the insurance companies, the banks, the

[19] These paragraphs are drawn with revisions from my foreword to the second edition of Seabright (2010), *The Company of Strangers*.

stock market, the government? Our civilization has been running smoothly – with some serious disruptions – for thousands of years, growing in complexity and power. Could it break down? Yes, it could, and to whom could we then turn to help us get back on the road? You can't buy a new civilization if yours collapses, so we had better keep the civilization we have running in good repair. Who, though, are the reliable mechanics? The politicians, the judges, the bankers, the industrialists, the journalists, the professors – the leaders of our society, in short – are much more like the average motorist than you might like to think: doing their local bit to steer their part of the whole contraption, while blissfully ignorant of the complexities on which the whole system depends. According to the economist and evolutionary thinker Paul Seabright (2010), the optimistic tunnel vision with which they operate is not a deplorable and correctable flaw in the system but an enabling condition. This distribution of partial comprehension is not optional. The edifices of social construction that shape our lives in so many regards depend on our myopic confidence that their structure is sound and needs no attention from us.

At one point, Seabright compares our civilization with a termite castle. Both are artifacts, marvels of ingenious design piled on ingenious design, towering over the supporting terrain, the work of vastly many individuals acting in concert. Both are thus by-products of the evolutionary processes that created and shaped those individuals, and in both cases, the design innovations that account for the remarkable resilience and efficiency observable were not the brainchildren of individuals, but happy outcomes of the largely unwitting, myopic endeavors of those individuals, over many generations. But there are profound differences as well. Human *cooperation* is a delicate and remarkable phenomenon, quite unlike the almost mindless cooperation of termites, and indeed quite unprecedented in the natural world, a unique feature with a unique ancestry in evolution. It depends, as we have seen, on our ability to engage each other within the "space of reasons," as Wilfrid Sellars put it. Cooperation depends, Seabright argues, on trust, a sort of almost invisible social glue that makes possible both great and terrible projects, and this trust is not, in fact, a "natural instinct" hard-wired by evolution into our brains. It is much too recent for that.[20] Trust is a by-product of social conditions that are

[20] Seabright points out that no band of chimpanzees or bonobos could tolerate the company of strangers – proximity to conspecifics who are not family or group members – that we experience with equanimity virtually every day, a profound difference. The (relative) calm with which many ungulate species can crowd together at a watering hole is not trust; it is instinctual indifference to familiar nonpredators, more like our attitude toward trees and bushes than our attitude toward other human beings in the landscape. Trust is a cultural phenomenon, as I observed in Dennett, 2017: chapter 7.

at once its enabling condition and its most important product. We have bootstrapped ourselves into the heady altitudes of modern civilization, and our natural emotions and other instinctual responses do not always serve our new circumstances.

Civilization is a work in progress, and we abandon our attempt to understand it at our peril. Think of the termite castle. We human observers can appreciate its excellence and its complexity in ways that are quite beyond the nervous systems of its inhabitants. We can also aspire to achieving a similarly Olympian perspective on our own artifactual world, a feat only human beings could imagine. If we do not succeed, we risk dismantling our precious creations in spite of our best intentions. Evolution in two realms, genetic and cultural, has created in us the capacity to know ourselves. But in spite of several millennia of ever-expanding intelligent design, we still are just staying afloat in a flood of puzzles and problems, many of them created by our own efforts of comprehension, and there are dangers that could cut short our quest before we – or our descendants – can satisfy our ravenous curiosity.

5. Home at last

This completes our journey from bacteria to Bach and back. It has been a long and complicated trek through difficult terrain, encountering regions seldom traveled by philosophers, and other regions beset by philosophers and typically shunned by scientists. I have invited you to take on board some distinctly counterintuitive ideas and tried to show you how they illuminate the journey. I would now like to provide a summary of the chief landmarks and remind you of why I found them necessary waypoints on the path. We began with the problem of the mind and Descartes's potent polarization of the issues. On one side, the sciences of matter and motion and energy and their support, thanks to evolution, of life; on the other side, the intimately familiar but at the same time utterly mysterious and private phenomena of consciousness. How can this dualist wound be healed? The first step in solving this problem, I argued, is Darwin's *strange inversion of reasoning*, the revolutionary insight that all the design in the biosphere can be, must ultimately be, the product of blind, uncomprehending, purposeless processes of natural selection. No longer do we have to see Mind as the Cause of everything else.

Evolution by natural selection can mindlessly uncover the *reasons without reasoners*, the free-floating rationales that explain why the parts of living things are arranged as they are, answering both questions: *How come?* and *What for?* Darwin provided the first great instance of *competence without comprehension* in the process of natural selection itself. Then *Turing's strange inversion* of reasoning provided an example, and a workbench for exploring

the possibilities, of another variety of competence without comprehension: computers, which unlike the human agents for which they were named, do not have to understand the techniques they exploit so competently. There is so much that can be accomplished by competence with scant comprehension – think of termite castles and stotting antelopes – that we are faced with a new puzzle: What is comprehension for, and how could a human mind like Bach's or Gaudí's arise? Looking more closely at how computers are designed to use information to accomplish tasks heretofore reserved for comprehending human thinkers helped clarify the distinction between "bottom-up" design processes exhibited by termites – and by natural selection itself – and "top-down" intelligent design processes. This led to the idea of *information as design worth stealing*, or buying or copying in any case. Shannon's excellent theory of information clarifies the basic idea – *a difference that makes a difference* – and provides it with a sound theoretical home, and ways of measuring information, but we need to look further afield to see why such differences are so valuable, so worth measuring in the first place.

The various processes of Darwinian evolution are not all the same, and some are "more Darwinian" than other processes that are just as real, and just as important in their own niches, so it is important to be a *Darwinian about Darwinism*. Godfrey-Smith's Darwinian Spaces is a good thinking tool for helping us plot not only the relations between the way different species evolve, but also the way evolution itself evolves, with some lineages exhibiting de-Darwinization over time.

Returning to the puzzle about how brains made of billions of neurons without any top-down control system could ever develop into human-style minds, we explored the prospect of decentralized, distributed control by neurons equipped to fend for themselves, including as one possibility *feral neurons*, released from their previous role as docile, domesticated servants under the selection pressure created by a new environmental feature: cultural invaders. *Words striving to reproduce*, and other memes, would provoke adaptations, such as revisions in brain structure in coevolutionary response. Once cultural transmission was secured as the chief behavioral innovation of our species, it not only triggered important changes in neural architecture but also added novelty to the environment – in the form of thousands of Gibsonian affordances – that enriched the ontologies of human beings and provided in turn further selection pressure in favor of adaptations – thinking tools – for keeping track of all these new opportunities. *Cultural evolution itself evolved* away from undirected or "random" searches toward more effective design processes, foresighted and purposeful and dependent on the comprehension of agents: intelligent designers. For human comprehension, a huge array of thinking tools is required. Cultural evolution de-Darwinized itself with its own fruits. This

vantage point lets us see the manifest image, in Wilfrid Sellars's useful terminology, as a special kind of artifact, partly genetically designed and partly culturally designed, a particularly effective *user-illusion* for helping time-pressured organisms move adroitly through life, availing themselves of (over)simplifications that create an image of the *world we live* in that is somewhat in tension with the scientific image to which we must revert in order to explain the emergence of the manifest image. Here we encounter yet another revolutionary *inversion of reasoning*, in David Hume's account of our knowledge of causation. We can then see human *consciousness as a user-illusion*, not rendered in the Cartesian Theater (which does not exist) but constituted by the representational activities of the brain coupled with the appropriate reactions to those activities ("and then what happens?").

This closes the gap, the Cartesian wound, but only a sketch of this all-important unification is clear at this time. The sketch has enough detail, however, to reveal that human minds, however intelligent and comprehending, are not the most powerful imaginable cognitive systems, and our intelligent designers have now made dramatic progress in creating machine learning systems that use bottom-up processes to demonstrate once again the truth of Orgel's Second Rule: Evolution is cleverer than you are. Once we appreciate the universality of the Darwinian perspective, we realize that our current state, both individually and as societies, is both imperfect and impermanent. We may well someday return the planet to our bacterial cousins and their modest, bottom-up styles of design improvement. Or we may continue to thrive, in an environment we have created with the help of artifacts that do most of the heavy cognitive lifting their own way, in an age of post-intelligent design. There is not just coevolution between memes and genes; there is codependence between our minds' top-down reasoning abilities and the bottom-up uncomprehending talents of our animal brains. And if our future follows the trajectory of our past – something that is partly in our control – our artificial intelligences will continue to be dependent on us even as we become more warily dependent on them.

References

Arnold, Frances (2013) *Frances Arnold Research Group*. Available at http://www.che.caltech .edu/groups/fha/Projects3b.htm.

Asadia, Ehsan et al. (2014) "Multi-Objective Optimization for Building Retrofit: A Model Using Genetic Algorithm and Artificial Neural Network and an Application", *Energy and Buildings*, 81: 444–456.

Bostrom, Nick (2014) *Superintelligence: Paths, Dangers, Strategies*, New York: Oxford University Press.

Chalmers, David (2010) "The Singularity: A Philosophical Analysis" *Journal of Consciousness Studies*, 17 (9–10): 7–65.

Chomsky, Noam (1975) *Reflections on Language*, New York: Pantheon Books.

Chomsky, Noam (2014) "Mysteries and Problems." October 18, https://www.youtube.com /watch?v=G8G2QUK_1Wg.

Chou, Christine et al. (2012) "Using Interactive Evolutionary Computation (IEC) with Validated Surrogate Fitness Functions for Redistricting", Presented at the *Genetic and Evolutionary Computation*, ACM: Philadelphia.

Cope, David and Hofstadter, Douglas R. (2001) *Virtual Music: Computer Synthesis of Musical Style*, Cambridge, Mass.: MIT Press.

Debner, James A. and Jacoby, Larry L. (1994) "Unconscious Perception: Attention, Awareness, and Control" *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 20 (2): 304–317.

Dehaene, Stanislas and Naccache, Lionel (2001) "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework", *Cognition*, 79 (1–2): 1–37.

Dennett, Daniel (1984) "A Route to Intelligence: Oversimplify and Self-monitor", Available at http://ase.tufts.edu/cogstud/papers/oversimplify.pdf.

Dennett, Daniel (1985) "Can Machines Think?" In *How We Know* (M. Shafto, ed.), San Francisco: Harper & Row.

Dennett, Daniel (1987) "Fast Thinking" in *The Intentional Stance*, Cambridge, Mass.: MIT Press.

Dennett, Daniel (1998) "Information, Technology, and the Virtues of Ignorance" in *Brainchildren: Essays on Designing Minds*, Cambridge, Mass.: MIT Press.
Dennett, Daniel (2001) "Are We Explaining Consciousness Yet?", *Cognition*, 79: 221–237.

Dennett, Daniel (2006) *Breaking the Spell: Religion as a Natural Phenomenon*, New York: Viking.

Dennett, Daniel (2012) "The Mystery of David Chalmers", *Journal of Consciousness Studies*, 19 (1–2): 86–95.

Dennett, Daniel (2013) "The Chinese Room" In *Intuition Pumps and Other Tools for Thinking*, New York: W. W. Norton.

Dennett, Daniel (2015) "The Singularity—An Urban Legend?" In *What to Think about Machines That Think* (John Brockman, ed.), New York: HarperCollins, pp. 85–88.

Dennett, Daniel (2017) *From Bacteria to Bach and Back: The Evolution of Minds*, New York, London: W.W. Norton & Company.

Descartes, René (1637) [1956], *Discourse on Method*, New York: Liberal Arts Press.

Domingos, Pedro (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, New York: Basic Books.

Edge World Question (2015) "What do you Think about Machines that Think", Available at https://www.edge.org/contributors/q2015.

Guston, Philip (2011) *Philip Guston: Collected Writings, Lectures, and Conversations* (Clark Coolidge, ed.), Berkeley, Los Angeles, London: University of California Press.

Hofstadter, Douglas (1985) *Metamagical Themas: Questing for the Essence of Mind and Pattern*, New York: Basic Books.

Hofstadter, Douglas (1995) *Fluid Concepts and Creative Analogies*, New York: Basic Books.

Hofstadter, Douglas and Sander, Emmanuel (2013) *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*, New York: Basic Books.

Katchadourian, Raffi (2015) "The Doomsday Invention: Will Artificial Intelligence Bring Us Utopia or Destruction?" in *New Yorker*, November 23, 64–79.

Kurzweil, Ray (2005) *The Singularity Is Near: When Humans Transcend Biology*, New York: Viking.

Littman, Michael L., et al. (1998) "Automatic Cross-Language Information Retrieval Using Latent Semantic Indexing" In *Cross-Language Information Retrieval*, New York: Springer, pp. 51–62.

McGinn, Colin (1991) *The Problem of Consciousness: Essays towards a Resolution*, Cambridge, Mass.: Blackwell.

Merikle, Philip M., Smilek, Daniel and Eastwood, John D. (2001) "Perception without Awareness: Perspectives from Cognitive Psychology", *Cognition*, 79 (1/2): 115–134.

Miller, George A., Galanter, Eugene and Pribram, Karl H. (1960) *Plans and the Structure of Behavior*, New York: Henry Holt.

Moravec, Hans P. (1988) *Mind Children: The Future of Robot and Human Intelligence*, Cambridge, Mass.: Harvard University Press.

Pugh, George Edgin (1978) *The Biological Origin of Human Values*, New York: Basic Books.

Rehder, Bob et al. (1998) "Using Latent Semantic Analysis to Assess Knowledge: Some Technical Considerations", *Discourse Processes*, 25 (2/3): 337–354.

Seabright, Paul (2010) *The Company of Strangers: A Natural History of Economic Life*, Rev. ed., Princeton, N.J.: Princeton University Press.

Smith, Shannon D. and Merikle, Philip M. (1999) "Assessing the Duration of Memory for Information Perceived without Awareness", Poster presented at the *3rd Annual Meeting of the Association for the Scientific Study of Consciousness*, Canada.

Sober, Elliott and Wilson, David S. (1995) "Some Varieties of Greedy Ethical Reductionism", *DDI*, 467–481.

Specter, Michael (2015) "The Gene Hackers: The Promise of CRISPR Technology", *New Yorker*, Nov. 16: 52.

Sullivan-Fedock, John (2011) "Increasing the Effectiveness of Energy Wind Harvesting with CFD Simulation-Driven Evolutionary Computation", Presented at the *Seoul CTBUH 2011 World Conference*, CTBUH: Seoul, South Korea.

CONTRIBUTORS

Aníbal Monasterio Astobiza is a Basque Government Posdoctoral Researcher. His research lies at the intersection of the cognitive, biological and social sciences exploring their philosophical underpinnings. He is a member of international (and national) research projects such as for example, EXTEND and INBOTS (Horizon 2020). He explores topics such as roboethics, machine ethics, ethics of AI, data ethics, ethics of neurotechnology, human enhancement, space ethics, applied ethics, bioethics, philosophy of neuroscience, philosophy of psychiatry, altruism, cooperation or philosophy of network science and complexity. Aníbal graduated in Philosophy (Universidad de Deusto) and obtained his PhD in Cognitive Science and Humanities at the Universidad del País Vasco/Euskal Herriko Unibertsitatea with a dissertation on social cognition. Monasterio Astobiza A. et al. (2019), Bringing inclusivity to robotics with INBOTS. Nature Machine Intelligence 1, 164.

Ben Goertzel is the CEO of the decentralized AI network SingularityNET, a blockchain-based AI platform company, and the Chief Science Advisor of Hanson Robotics, where for several years he led the team developing the AI software for the Sophia robot. Dr. Goertzel also serves as Chairman of the Artificial General Intelligence Society, the OpenCog Foundation, the Decentralized AI Alliance and the futurist nonprofit Humanity+. Dr. Goertzel is one of the world's foremost experts in Artificial General Intelligence, a subfield of AI oriented toward creating thinking machines with general cognitive capability at the human level and beyond. He also has decades of expertise applying AI to practical problems in areas ranging from natural language processing and data mining to robotics, video gaming, national security and bioinformatics. He has published 20 scientific books and 140+ scientific research papers, and is the main architect and designer of the OpenCog system and associated design for human-level general intelligence.

Caterina Moruzzi is a postdoctoral researcher at the University of Nottingham where she obtained her PhD in 2018 with a thesis on the ontology of musical works. She has a MA in Philosophy at the Università di Bologna and she graduated in Piano performance from the Conservatorio G.B. Martini in Bologna. She is actively contributing to research in aesthetics and Philosophy of Art by presenting at conferences worldwide and by publishing on high-impact journals. In her current research she investigates the impact of revolutionary development of AI technology on ontological, aesthetic, and evaluative considerations on music. In particular, she aims to answer two

research questions: 'How does the development of computer-generated music affect the ontology of musical works?' and 'Do we deem computer-generated music as creative as human-generated music? If not, why?'. The scope of this research is to fill the existing gap in the literature in regard to the impact that cutting-edge developments in AI have on aesthetic and ontological debates on music and art.

Cody Turner is a third year PhD student in philosophy at the University of Connecticut. His areas of research include the philosophy of consciousness, the philosophy of artificial intelligence, and applied ethics. He is a member of the AI, Mind, and Society Group at the University of Connecticut, and has a teaching appointment with the Computer Science & Engineering department. He currently has one other publication entitled 'Could you Merge with AI? Reflections on the Singularity and Radical Brain Enhancement' (with Susan Schneider, forthcoming in Oxford Handbook of Ethics of AI, Oxford University Press, 2020).

Daniel C. Dennett was educated at Harvard (BA 1963) and Oxford (DPhil, 1965) and taught at UC Irvine (1965-71) before moving to Tufts University, where he has taught ever since. He is currently University Professor and Austin B. Fletcher Professor of Philosophy. He is the author of many books, including "Consciousness Explained" (1991), "Darwin's Dangerous Idea" (1995) and "From Bacteria to Bach and Back" (2017) and hundreds of articles.

David Pearce is author of The Hedonistic Imperative (1995), which advocates using biotechnology to abolish suffering throughout the living world in favour of gradients of superhuman bliss. In 1998, Pearce co-founded with Nick Bostrom The World Transhumanist Association, now rebranded as Humanity Plus. Transhumanists urge the use of technology to create a "triple S" civilisation of superhappiness, superlongevity and superintelligence.

Eray Özkural has obtained his PhD in computer engineering from Bilkent University, Ankara. He has a deep and long-running interest in human-level AI. His name appears in the acknowledgements of Marvin Minsky's The Emotion Machine. He has collaborated briefly with the founder of algorithmic information theory Ray Solomonoff, and in response to a challenge he posed, invented Heuristic Algorithmic Memory, which is a long-term memory design for general-purpose machine learning. Some other researchers have been inspired by HAM and call the approach "Bayesian Program Learning". He has designed a next-generation general-purpose machine learning architecture. He is the recipient of 2015 Kurzweil Best AGI Idea Award for his theoretical contribution to universal induction. He has previously invented an FPGA virtualization scheme for Global Supercomputing, Inc. which was internationally patented. He has also proposed a cryptocurrency called

Cypher, and an energy based currency which can drive green energy proliferation. You may find his blog at http://log.examachine.net and some of his free software projects at https://github.com/examachine.

Federico Pistono is an Entrepreneur, Angel Investor, Researcher, and Educator. He's author of the international success "Robots Will Steal Your Job, But That's OK", has co-founded four startups, he's Futurist and Technology Expert in the TV show "Codice: la vita è digitale" (Code: life is digital), which airs on Rai1, and is Former Head of Blockchain of HyperloopTransportation Technologies. He's founder and CEO of Exponential Thinking, which invests in early stage startups, and creator ofMavericks, a school for Entrepreneurs, Investors, and Innovators. He holds a BSc in Computer Science from the University of Verona, and in 2012 he graduated from Singularity University, NASA Ames Research Park.

Frederic Gilbert currently research consists in exploring the ethics of novel implantable brain-computer interfaces operated by Artificial Intelligence. At the time of writing this bio, he is an Australian Research Council (ARC) Discovery Early Career Research Award recipient and Senior Lecturer in Ethics at CALE, School of Humanities, at the University of Tasmania. He is affiliated with the Ethics, Policy & Public Engagement program of the ARC Australian Centre of Excellence for Electromaterials Science (ACES), and concomitantly an Ethics Consultant for the Centre for Neurotechnology, for which he conducts research at the University of Washington, in Seattle, USA. He has published over 65 articles in bioethics and neuroethics. Some of his main publications related to this chapter are: Gilbert F., O'Brien T., Cook, M. (2018) "The Effects of Closed-Loop Brain Implants on Autonomy and Deliberation: What are the Risks of Being Kept in the Loop?" In Cambridge Quarterly of Healthcare Ethics, pp. 1-12; Gilbert F., Cook, M., O'Brien, T., Illes J. (2019) "Embodiment and estrangement: Results from a first-in-human intelligent BCI trial" in Science and Engineering Ethics, 25 (1): 83-96; Gilbert, F. (2015) "A threat to Autonomy? The Intrusion of Predictive Brain Devices". American Journal of Bioethics Neuroscience, 6 (4): 4-11.

Gabriel Axel Montes, Ph.D. Candidate, is a neuroscientist, educator, and musician. His 12+ years of research in cognitive neurobiology and mind-body practice has elucidated both endogenous and technological mechanisms for intervening in the brain's process of modelling reality and perception. His recent research focuses on informing the development of beneficial artificial general intelligence (AGI) with insights from the non-ordinary consciousness (NOC) of mind-body practices. Gabriel is known for fusing insights from rigorous NOC self-practice into the design of AGI, and virtual/augmented reality and humanitarian projects involving resilience education. Gabriel is an international educator on the intersection of neuroscience and mind-body practice. As

Founder of Neural Axis®, he leads and consults on neuroscience-based technology design, strategy, and behavioural change. He leads organisations, communities, and individuals in applying neuroscience to first-person experience. For more about Gabriel, please visit gabrielaxel.com.

Hasse Hämläinen received his PhD in Philosophy from the University of Edinburgh (2015). At the time of writing, he is working as a QA analyst for a leading IT company and as a project researcher at the Jagiellonian University in Kraków. His research is focused on the history of moral theory, especially on ancient and the 18th century moral thinking, and, recently, also on charting the ethical challenges and limits of AI technologies. His publications include a co-edited monograph, The Sources of Secularism: Enlightenment and Beyond (Palgrave Macmillan, 2017) and his work has been published, for example, in International Philosophical Quarterly, Diametros, Journal of Ancient Philosophy and in The Philosophical Forum.

Kulvinder Panesar is a strategically focused senior computing professional and an academic for over twenty years. Her current role is a Senior Lecturer of Computer Science at the School of Art, Design and Computer Science at York St John University, York in United Kingdom. Her research interest is NLP (Natural Language Processing) in AI (Artificial Intelligence), and meaning and knowledge representation in conversational software agents (CSAs), and more recently, Chatbots and conversational AI. This research area is multi-disciplinary spanning agent thinking, linguistics, and computational linguistics, natural language processing (NLP), knowledge representation (KR), Semantic Web (SW), artificial intelligence (AI) and data science. Current teaching commitments include undergraduate and postgraduate subjects in: computer science, software engineering design and development, website design and development, advanced database and networks; programming, professional practice; project management, project supervision, mathematics for computer science, philosophies of technology, and artificial intelligence. Previous teaching undertaken include: management information systems, marketing and e-commerce strategy (customer relationship management) and development, enterprise computing solutions, systems analysis and design, formal methods and research methods. Additional roles undertaken over the years are course director, module leader, final year project coordinator and supervisor, placement co-ordinator, 'study abroad' academic advisor, researcher, internal verifier, mentor, STEM ambassador, external assistant examiner (EdExcel), ICT teacher, IT support/programmer/trainer contractor, and systems analyst. She is a professional member (MBCS) of the British Computer Society (BCS), and a Fellow of the Higher Education Academy (AdvanceHE) and currently organise and co-ordinate the Women in Computer Science (WICS) group at York St John University.

Mariana Chinellato Ferreira is a Brazilian researcher, graduate at Social Communication with specialization in Advertising and Publicity from the Catholic University of Santos (2002) and graduate at Modern Languages, major in Portuguese and English languages and Literature from the Catholic University of Santos (2009). Has an exchange program period at the University of Coimbra (Portugal - 2007/2008) in Modern Languages. Has professional experience in Linguistics, focusing on Applied Linguistics, acting on the following subjects: Modern Foreign Language (English), English for Specific Purposes, Portuguese as a Second Language, Translation (Portuguese – English) and Education. Professional academic experience in English and American Literature, Portuguese and Brazilian Literature, Theory of Literature, Theater and Poetry. She has a Master's degree in Architecture at University of São Paulo - São Carlos with the project: "City and Literary form: Urban representation in the Brazilian contemporary literature". Currently, she is a PhD Candidate of the PhD Program in Materialities of Literature at the University of Coimbra with a research about questions of language and creativity in Narrative Generation Systems in a partnership with joint supervision of the Centre for Informatics and Systems of the University of Coimbra where she develops a new model of narrative in order to apply into a prototype system of automatic narrative generation in Portuguese language. Her last publish article discusses new models of narrative in computer-generated literature: "Seeking the Ideal Narrative Model for Computer-Generated Narratives", presented at the CC-NLG 2018, in Tilburg, the Netherlands.

Natasha Vita-More is an author, humanitarian, and innovator whose work focuses on longevity and regenerative generations. As a motivational speaker, she focuses on causes and solutions, while fostering meaningful acknowledgement of others' works who have aspired to identify human potential. She is called "An early adapter of revolutionary changes" (Wired, 2000) and "Advocates the ethical use of technology to expand human capacities" (Politico, 2017). Natasha was the Lead Scientific Researcher on the Memory Project, which scientific breakthrough concerns long-term memory of C. elegans and cryonics (2015). As a proponent for mitigating aging, Natasha introduced the seminal field of Human Enhancement for longevity in academics. Her expertise in the field of ethics has produced high-level scholarship for undergrad and graduate students in the fields of computer science, cybersecurity, robotics, gaming, and business fields. As an entrepreneur, her experience within the domain of foresight studies has established principles and practices for assessing humanity's potential futures. Natasha originated the original "Whole-Body prosthetic", a seminal innovation comprised of nanomedicine, AI, and robotics that spearheaded alternatives to

biological aging and received international recognition. Currently, she is a Professor of Humanities in Ethics, Business, and Entrepreneurship, and serves as the Executive Director of Humanity+. She is also a Fellow at the Institute for Ethics and Emerging Technologies, and author of numerous academic articles and books. Some publications: "Persistence of Long-Term Memory in Vitrified and Revived Caenorhabditis elegans" or "The Transhumanist Reader". More can be find here: https://natashavita-more.com/publications/.

Nicole Hall most recently held a Fernand Braudel research fellowship at the Institut Jean Nicod (Ecole Normale Supérieure) and taught aesthetics at the Université Paris 8. Her PhD, on the nature of aesthetic experience, is from the University of Edinburgh. Her research is at the intersection of aesthetics, environmental philosophy, the philosophy of mind and consciousness, and cognitive and neuroscience. She has a forthcoming publication on Adam Smith's aesthetic psychology and is working on the role of perception in aesthetic experience.

Peter DePergola II is Director of Clinical Ethics at Baystate Health, Assistant Professor of Medicine at University of Massachusetts Medical School, and Assistant Professor of Bioethics and Medical Humanities at Elms College, where he serves concurrently as Director of the Center for Ethics, Religion, and Culture. Dr. DePergola earned his B.A. degree in philosophy and religious studies at Elms College, his M.T.S. degree in ethics at Boston College, and his Ph.D. degree in healthcare ethics at Duquesne University. He completed his residency in neuroethics at University of Pittsburgh School of Medicine, his fellowship in neuropsychiatric ethics at Tufts University School of Medicine, and his advanced training in neurothanatological ethics at Harvard Medical School. Dr. DePergola's recent book, Forget Me Not: The Neuroethical Case Against Memory Manipulation(Vernon Press, 2018), has been critically acclaimed as a landmark achievement in the field of neuroethics, and was the #1 New Release in Medical Ethics on Amazon.com at the time of its publication. The recipient of numerous awards for academic scholarship and healthcare leadership, his current research explores the theoretical and empirical metaphysics of hope, particularly as it relates to the neuroethics of moral virtue and end-of-life decision making.

René Mogensen holds a PhD from Birmingham City University/Royal Birmingham Conservatoire in the UK, an MM from the Royal academy of Music in Aarhus, Denmark, an MA from New York University and a BA from the University of Rochester, NY, USA. His research in music, computer music, and artificial intelligence in music has recently been published in books by Springer, Éditions Delatour, and as articles in the Journal of Creative Music Systems, the Early Music journal, as well as in international conference

proceedings (CSMC, AISB, EuroMAC). He is the recipient of grants from the Danish Arts Foundation, Meet the Composer USA, John Anson Kittredge Fund, Yvar Mikhashoff Trust for New Music, and other funders. His numerous original music works for ensembles and soloists (with and without computers) have been performed around Europe, the USA and Asia. He has participated as a performer/composer in many music concerts during the past three decades around the USA and Europe. He currently teaches undergraduates and postgraduates in areas of creative music technology as well as in music improvisation and composition at the Royal Birmingham Conservatoire in the UK.

Ricardo Morte Ferrer is a lawyer and Data Protection Consultant (https://sdm2.de/), he is also a certified ITIL-F and Lead Auditor ISO 27001. He is a member of the Working Group Digitalization and Health at the German Academy for Ethics in the Medicine. PhD candidate at the University of Granada working on"Ethical and Philosophical Aspects of Privacy". Nowadays is researcher in INBOTS CSA network: Inclusive Robotics for a better Society [www.inbots.eu/], EXTEND: Bi-directional Hyper-connected Neural System [www.extend-project.eu/] and member of the core team at GNU Health.

Roman V. Yampolskiy is a Tenured Associate Professor in the department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville. He is the founding and current director of the Cyber Security Lab and an author of many books including Artificial Superintelligence: a Futuristic Approach. During his tenure at UofL, Dr. Yampolskiy has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and KAS Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a Senior member of IEEE and AGI, and former Research Advisor for MIRI and Associate of GCRI. Roman Yampolskiy holds a PhD degree from the Department of Computer Science and Engineering at the University at Buffalo. He was a recipient of a four year NSF (National Science Foundation) IGERT fellowship. Before beginning his doctoral studies Dr. Yampolskiy received a BS/MS (High Honors) combined degree in Computer Science from Rochester Institute of Technology, NY, USA. After completing his PhD dissertation Dr. Yampolskiy held a position of an Affiliate Academic at the Center for Advanced Spatial Analysis, University of London, College of London. Dr. Yampolskiy is an alumnus of Singularity University (GSP2012) and a Visiting Fellow of the Singularity Institute (Machine Intelligence Research Institute). Dr. Yampolskiy's main areas of interest are AI Safety, Artificial Intelligence, Behavioral Biometrics, Cybersecurity, and Genetic Algorithms.

Soenke Ziesche holds a PhD in Natural Sciences from the University of Hamburg, which he received within a Doctoral Program in Artificial Intelligence. He has worked since 2000 for the United Nations in the humanitarian, recovery and sustainable development fields as well as in data and information management. In addition to UN Headquarters in New York he has been posted to Palestine, Sri Lanka, Pakistan, Sudan, Libya, where he was temporarily the highest UN representative during the revolution, South Sudan, Bangladesh and Maldives. He is currently a consultant, mostly for the UN, for Disaster Information Management as well as AI. In this capacity he has served, on behalf of the UN, the Governments of Bangladesh, Maldives and Sri Lanka respectively. In 2017 he produced a 137-pages UN report on opportunities for big data, AI and emerging technologies in the implementation and review of the UN Sustainable Development Goals. His latest academic publications on AI safety are "AI & Global Governance: A Seat at the Negotiating Table for AI? Opportunities and Risks", "Potential Synergies Between The United Nations Sustainable Development Goals And The Value Loading Problem In Artificial Intelligence" and, together with Roman V. Yampolskiy, "Towards AI Welfare Science and Policies".

Stephen Rainey is a research fellow in the Oxford Uehiro Centre for Practical Ethics. He is a working in the Horizon2020-funded project BrainCom which is developing therapeutic brain-computer interfaces that will enable communication for users with debilitating speech conditions. Dr Rainey studied philosophy in Queen's University Belfast and obtained his PhD in 2008 with a thesis on rationality. He has taught a range of philosophical topics, and worked on a number of European Commission-funded research projects. These have included work on ethics, emerging technologies, and governance. He contributes to European Commission ethical and scientific evaluation panels for the funding of research projects. These have included proposals for ERC and Marie-Curie grants. Stephen has research interests in the philosophy of language, rationality, governance, and artificial intelligence.

Stevan Harnad's research interest is in cognitive science, open access and animal rights. He is currently professor of psychology at Université du Québec à Montréal (UQAM) and professor emeritus of cognitive science at the University of Southampton. Elected external member of the Hungarian Academy of Sciences in 2001 (resigned in protest, 8 October 2016), he was founder and editor-in-chief of the journal Behavioral and Brain Sciences 1978-2002 and Canada Research Chair in cognitive science 2001-2015. His research is on categorization, communication, cognition, and consciousness. He has written on categorical perception, symbol grounding, origin of language, lateralization, the Turing test, distributed cognition, scientometrics, and animal sentience. He is founder and editor-in-chief of the journal Animal Sentience.

Steven S. Gouveia is finishing his PhD at the University of Minho (Portugal), under the supervision of the philosopher Manuel Curado (University of Minho) and the neuroscientist Georg Northoff (University of Ottawa), funded by the Science and Technology Foundation. His primary focus of research is on the relationship between Neuroscience and Philosophy. He is a visiting researcher at the Minds, Brain Imaging and Neuroethics Group at the Royal Institute of Mental Health, Uni. Ottawa. He is a researcher of the Lisbon Mind & Reasoning Group (Nova University of Lisbon) and the Mind, Language and Action Group (Uni. Porto). He published 8 academic books e.g. (2018) "Perception, Cognition and Aesthetics" (eds.) (Routledge) or "Automata's Inner Movie: Science and Philosophy of Mind" (eds.) (Vernon Press) and, forthcoming, "The Science and Philosophy of Predictive Processing (eds.) (Bloomsbury). He is the host of an international documentary on AI with the participation of international researchers such as Peter Singer and Paul Thagard. He is finishing writing is new book "Homo Ignarus: Rational Ethics for an Irrational World". He was one of the speakers at the famous "Science of Consciousness Conference" (2019) and published several articles on AI, mind and brain, ethics and aesthetics. More: stevensgouveia.weebly.com.

Tomislav Miletić, born on May 7, 1984, in the city of Rijeka, Croatia. My personal and academic interests predominantly lie in exploring the ethical and social impacts of Artificial Intelligence inside the paradigm of Human Enhancement. I am especially focused on the prospect of symbiotic human-Ai systems, their hybrid moral-epistemic agency, status and system architecture and the relevant ethical questions concerning their design and use. Currently engaged in finishing my doctoral thesis on the subject of "Human-AI symbiosis: the possibility of moral augmentation". Relevant publications: Miletić T. "Extraterrestrial artificial intelligences and humanity's cosmic future: Answering the Fermi paradox through the construction of a Bracewell-Von Neumann AGI" in Journal of Evolution and Technology, Vol. 25:1, 2015, 56-73 or Miletić T. "Human Becoming: Cognitive and Moral Enhancement Inside the Imago Dei Narrative" in Theology And Science, Vol. 13:4, 2015, 425-445.

Txetxu Ausín is a Tenured Scientist at the Institute for Philosophy, CSIC (Spanish National Research Council) [www.ifs.csic.es]. He is the Director of the Applied Ethics Group (GEA). PhD in Philosophy (University of the Basque Country) and First Prize of the Year 2000. Invited Professor in several universities and researcher at the Institute for Democratic Governance Globernance. His research areas involve public ethics, bioethics, deontic logic, human rights, and philosophy of robotics and ICT. [Member of OECD Expert Group on Research ethics and new forms of data for social and economic research]. Editor and author of publications about this issues, is the Editor in Chief of the international

journal DILEMATA on applied ethics [www.dilemata.net]. Nowadays is researcher in INBOTS CSA network: Inclusive Robotics for a better Society [www.inbots.eu/], EXTEND: Bi-directional Hyper-connected Neural System [www.extend-project.eu/] and BIODAT: Datos en salud[umucebes.es/]. Research Vice-President of ASAP-Spain (Academics Stand Against Poverty) and President of the Spanish Network of Philosophy [http://redfilosofia.es/]. Independent Member of the Public Ethics Commission of the Basque Government and Patron of the Foundation Cluster Ethics of the Basque Country and of Legal and Social Studies (EJYS Foundation), he also collaborates in the Committee of Experts on Aging of the Fundación General CSIC. Some publications: "Exclusion from healthcare in Spain: The responsibility for omission of due care". In Gaisbauer, Schweiger, Sedmak (eds.), Ethical Issues in Poverty Alleviation. Dordrecht: Springer, 2016. ; "Research Ethics and New Forms of Data for Social and Economic Research". OECD Science, Technology and Industry Policy Papers, No. 34, OECD Publishing, Paris.

Vernor Vinge has won five Hugo Awards, including one for each of his last three novels, A Fire Upon the Deep (1992), A Deepness in the Sky (1999), and Rainbow's End (2006). Known for his rigorous hard-science approach to his science fiction, he became an iconic figure among cybernetic scientists with the publication in 1981 of his novella "True Names," which is considered a seminal, visionary work of Internet fiction. His many books also include Marooned in Realtime and The Peace War. Born in Waukesha, Wisconsin and raised in Central Michigan, Vinge is the son of geographers. Fascinated by science and particularly computers from an early age, he has a Ph.D. in computer science, and taught mathematics and computer science at San Diego State University for thirty years. He has gained a great deal of attention both here and abroad for his theory of the coming machine intelligence Singularity. Sought widely as a speaker to both business and scientific groups, he lives in San Diego, California.

INDEX

A

B

Daniel C. Dennett, Tufts University, USA

C

D

E

F

G

H

I

J

K

L

M

N

O

P

R

S

T

U

W

Y

Z