S.M. Amadae, "Life as Algorithm," in Futures, ed. by Jenny Andersson and Sandra Kemp, Oxford UP, 2021, pp. 462-483.

OXFORD
TWENTY-FIRST CENTURY
APPROACHES TO LITERATURE

Futures

Text

Edited by

JENNY ANDERSSON
AND
SANDRA KEMP

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2021
Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2020934982

ISBN 978–0–19–880682–0

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

CONTENTS

III FUTURES AS SALVATION AND APOCALYPSE

IV FUTURES OF LIFE

V FUTURE WORLDS

27

LIFE AS ALGORITHM

S. M. AMADAE

> Brains may be regarded as analogous in function to computers... Every decision that a survival machine takes is a gamble, and it is the business of genes to program brains in advance so that on average they take decisions that pay off. The currency used in the casino of evolution is survival, strictly gene survival, but for many purposes individual survival is a reasonable approximation.[1]

> First we humans discovered how to replicate some natural processes with machines, making our own wind and lightning, and our own mechanical horsepower. Gradually, we started realizing that our bodies were also machines. Then the discovery of nerve cells started blurring the borderline between body and mind. Then we started building machines that could outperform not only our muscles, but our minds as well. So in parallel with discovering what we are, are we inevitably making ourselves obsolete?[2]

> [W]e make decisions all the time, and... every decision we make reveals something about our goals [aka preferences]... The hope is therefore that by observing lots of people in lots of situations (either for real or in movies and books), the AI can eventually build an accurate model of all our preferences [and construct our environments accordingly].[3]

The emerging and even triumphant view of life as an algorithm stands in relief against a modern background in which life, celebrated as manifesting animism, was contrasted to mechanism. Even René Descartes and Immanuel Kant, who celebrated the key roles of matter and mechanism in existence, recognized an important role for mind in writing the narrative of events conducted by humans as agents. Sentient awareness can be a prime mover of the will and, according to Kant, can initiate a

[1] Richard Dawkins, *The Selfish Gene* (Oxford University Press, 1976), 49, 55.
[2] Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (London: Penguin, 2017), 319–20.
[3] Tegmark, *Life 3.0*, 262.

causal chain of events.[4] Since the late 1960s, there has been a shift in academic expertise on the nature of life and intelligence from naturalists such as Charles Darwin and William James, and biologists including James D. Watson and Francis Crick, to mathematical modellers and specialists in computation. Thus physicists, game theorists, and economists now take a leading role in explaining what life is, and in offering visions—some aspirational—of human life's potential futures.[5] Although life remains a mystery, in the late twentieth century and early twenty-first century we can identify a new and prevalent strand of thought engaging the nature and purpose of biological existence. Theorists in this new tradition recognize that digital and information technologies represent a fourth revolution of human understanding which further displaces human beings from the centre of God's cosmos, the apex of creation, and the singularity of rational unified agency.[6] This essay considers the writings of Richard Dawkins and Max Tegmark. The first was inspired by game theory as the means to study evolution, and the second assesses how the materialist basis of life could place humans and artificial intelligence on the same footing. Both thinkers consider life as an algorithm programmed to achieve success in survival and reproduction. Algorithms are material processes with the property that their physical structure dictates their outcomes. Once programmed into a physical substrate, they execute the same procedure every time they are executed. Algorithms leave no role for a form of agency that initiates its own causal chain as a function of, for example, sentient awareness and self-direction.[7]

This account of life as algorithm, or causal process not materially transformed by the contents of consciousness, has determinist implications and is currently mainstream. In this view, mind, feeling, and meaning are derivative of the causal processes from which they arise. The physical processes giving rise to life, and phenomenal experience, are deterministic and stochastic, in keeping with Newtonian physics and quantum theory. One way to understand this position is to think of experiences and mental projections as shadows caused by material processes. However, for the time being this view is based on metaphysical argumentation.[8] Considered in accordance with the charity principle, this view of both life and intelligence as algorithmic process appears to be fully consistent with

[4] Immanuel Kant, *Kants gesammelte Schriften*, ed. Königlichen Preußischen (later Deutschen) Akademie der Wissenschaften (Berlin: Georg Reimer (later Walter De Gruyter), 1913), vol. 5, 96–8.
[5] Along these lines, consider the work of Herbert Gintis, *Bounds of Reason* (Princeton University Press, 2009) and Ken Binmore, *Natural Justice* (Oxford University Press, 2005).
[6] See, for example, Luciano Floridi, *Fourth Revolution: How the Infosphere is Reshaping Human Reality* (Oxford University Press, 2014), 87–100.
[7] See, for example, Immanuel Kant, *Groundwork of the Metaphysic of Morals*, ed. and trans. H. J. Paton (New York: Harper and Row, 1964).
[8] For an attempt to ground this debate in natural scientific argumentation, see Scott Aaronson, 'The Ghost in the Quantum Turing Machine', in S. Barry Cooper and Andrew Hodge (eds.), *The Once and Future Turing: Computing the World* (Cambridge University Press, 2016), 193–296, https://arxiv.org/pdf/1306.0159.pdf.

the principles of natural science and the laws of physics. And yet, not only are there grounds for debate, but there are implications for how we imagine and thus potentially realize shared futures. The view of existence that is displaced by the materialist view provides a role for meaning, reflexive self-awareness, and intelligibility in human action and its capacity to realize ends.[9] According to the materialist conception, all possible futures simply reflect an arrangement of the physical building blocks of the universe. Opportunities are possible outcomes realizable given the current configuration of all physical entities. This approach replaces a view of aspirational and shared intention with a view of collective life as decomposable into individual-, or even genetic-, level competition. Actors satisfy preferences programmed into their physical structures to guide behaviour consistent with evolutionary goals. Neither individual consciousness, which may have holistic comprehension of situations, nor shared collective awareness, instantiates agency. All action is solely the product of physical processes, and reflexive awareness does not add a new dimension of intelligence or self-guidance, either individually or collectively.

Dawkins and Tegmark reflect two related perspectives on life which explain its nature and take positions on its significance and meaning. It is striking that these theorists are positioned as progressives in politics, and each of them, while advocating inclusive economic policies and an egalitarian respect for personhood and individualism, also proposes perspectives that potentially alienate religious conservatives and establishment traditionalists throughout the Judeo-Christian-Islamic West and Middle East.[10]

Evolutionary game theory, used by Dawkins, offers the view that life must solve objective optimization problems. For Dawkins, life is circumscribed by the imperative to survive and propagate, which is algorithmically programmed. Tegmark concedes that intelligence is optimization and that life solves optimization problems. Yet he puts forward the view that if humans can employ artificial intelligence in a constructive manner then, rather than decay in world of gross inequity, nuclear war, and technological destruction, life can master the control of universally available sources of energy and achieve goals without limit.[11]

In the lexicon offered by Tegmark, Dawkins's analysis remains confined to Life 2.0, while his own analysis explores how we may view artificial intelligence (AI) as coextensive with life extended as it were to a higher level of capability for optimization and problem-solving. Tegmark's novel insight is to offer that intelligence is computation: the execution of an algorithm which is embedded in deterministic cause–effect processes. Tegmark grounds phenomenal consciousness in computation which, by definition, lacks intelligible understanding of the significance of action. Meaning is

[9] John Searle, for example, who also works within naturalistic, evidence-based argumentation, *Rationality in Action* (Cambridge, MA: MIT Press, 2001).

[10] See the Wikipedia article on Richard Dawkins's political views to get an overview of how this vision of life reduced to the laws of physics has generated controversy, https://en.wikipedia.org/wiki/Political_views_of_Richard_Dawkins.

[11] Tegmark, *Life 3.0*, 260–1.

Figure 27.1 Optimistic vision of algorithmic life embraced by the Future of Life Institute. Image courtesy of i-stock.

ascribed after the fact of action and is not generative of agency. Tegmark looks to a bright possible future wherein AI assists humans in satisfying their preferences, and anticipates a hybrid world of human and artificial agents, with their boundaries blurred, as reflected in this image used on the Future of Life Institute's website (Figure 27.1).

Hardwired Selfishness

In addressing the topic of treating life as an algorithm, focusing on Richard Dawkins's work is useful because he wrote at the forefront of the movement treating life and intelligence as programmed. He distilled the significance of the research of evolutionary game theorists, and made his theories accessible to a large readership by popularizing his ideas.[12] In articulating the implications of this approach to understanding life, Dawkins argues that organisms' behaviour is programmed by their genetic codes which are selected for promoting the survival of individual members of species.[13] Dawkins has been so successful in proclaiming his views that his book *The Selfish Gene* (1976) is Oxford University Press's best-selling book of all time. He has also managed to invert our understanding of Charles Darwin's evolutionary

[12] Dawkins drew on the cutting edge of evolutionary game theory developed by John Maynard Smith, 'The Theory of Games and the Evolution of Animal Conflict', *Journal of Theoretical Biology* 47, no. 209 (1974), 209–21, and R. L. Trivers, 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology* 46, no. 1 (1971), 35–57.

[13] This is a one-sentence synopsis of Dawkins's argument running throughout *The Selfish Gene*; see, for example, the preface to the first edition.

theory as natural selection via a process of descent with modification to a view in which genes replicate over generations without modification and endlessly compete for survival in repeating games. According to Charles Darwin's *Origin of Species* (1859), life forms organized into species developed over prolonged periods of time during which their physical traits transformed because those favourable for survival as a response to environmental constraints were selected for and therefore were expressed in the next generation. Just as animal breeders could develop lineages and establish pedigrees to accentuate specific features as criteria for breeding, so natural selection could have the same effect, although without intentional design. For Darwin, life forms exist as interdependent species incessantly adapting to their environments. According to Dawkins, the key unit of biological organization is the gene, which he describes as fixed and never changing. This immutable gene is either perpetuated into the next generation or not, as a function of its ability to program action associated with survival.

Like Sigmund Freud, who suggested that much of human action is the product of subconscious drives and impulses outside of individuals' awareness and conscious control, Dawkins argues that the human intellect and sense of self-given purpose are superficial.[14] Ideation and meaning are a by-product of physical processes that must occur in order to sustain the existence of life's material basis; sentient awareness is not directly relevant to survival and reproduction. In the account he puts forward, genetic code in the form of segments of DNA must program behaviour that assists the macro-organism containing it to be biologically viable. Genes that prevail over competitors are represented in ensuing generations of organisms. Dawkins views genes as selfish because, according to his analysis, their survival depends on programming behaviour that is dedicated to their host's survival and reproduction on an individual, and not a group, basis.[15]

Dawkins's argument proceeds by analytic derivation, and not empirical study, and defers to ad hoc discussion of cases which seem to corroborate his formal model. His theory is entirely deduced using game theory, which provides a mathematical framework for modelling strategic competition. Another way to consider game theory is that it provides a method to solve multiple constraint problems in which every actor attempts to maximize an environmentally limited source of value in competition with like actors. Although originally designed to provide strategic solutions to parlour games with fixed rules linking actions to rewards, after its inventors John von Neumann and Oskar Morgenstern's 1944 publication of *Theory of Games and Economic Behavior*, successive generations of theorists put it to alternative uses, initially applying it to warfare.[16] It was first applied to model organisms confronting natural selection in evolution in the 1960s.[17] Dawkins relied on this novel research literature to construct his selfish gene argument.

Understanding Dawkins's position that 'genes are master programmers, and they are programming for their lives', and that the environmental pressures of natural selection require them to program selfish behaviour, is inseparable from learning the rudiments of game theory.[18] He introduces readers to games, or strategic interactions, which like parlour games have rules associating particular actions with given outcomes—pay-offs—depending on the actions every competitor takes. He leads readers to consider how we would proceed '[i]f we were to program a computer to simulate a model survival machine making decisions about whether to behave altruistically'.[19] Values need to be associated with behavioural choices. Dawkins promises 'a very over-simplified example', which relies on the concept of expected utility from game theory: 'I am an animal who has found a clump of eight mushrooms. After taking account of their nutritional value, and subtracting something for the slight risk that they might be poisonous, I estimate they are worth +6 units each.'[20] I need to decide whether to share the remaining mushrooms after I am satiated with next of kin or not. Before considering strategic calculations, I can perform a parametric calculation just considering outcomes without competitors. Next, strategic considerations enter into the calculation to determine whether to share. Throughout the game-theoretic account of strategic action, agents act *as if* they make conscious decisions. Dawkins clearly makes this point, observing, '[a]nimals have to be given by their genes a simple rule for action, a rule that does not involve all-wise cognition of the ultimate purpose of the action, but a rule that works nevertheless, at least in average conditions.'[21] Here Dawkins conveys the crucial point about computable algorithms that encompass strategic rationality. Those actors animated by algorithmic decision rules do not grasp the context or purpose of their action. Hence, they lack intelligible grasp of the meaning and implications of their conduct.

In order to definitively conclude that biological actors cannot be altruistic *and* survive the rigours of natural selection, Dawkins relies on a formal concept developed in game theory: the Evolutionarily Stable Strategy (ESS). This concept builds on the original mathematical game theory of the 1950s, specifically John F. Nash Jr's mutual-best-reply equilibrium. According to Nash's solution concept, an equilibrium of individuals' action choices signifies that from that outcome no single individual would choose an alternative action to achieve a preferred outcome,

[14] See, for example, Sigmund Freud, *Civilization and its Discontents* (New York: Norton, 1962).

[15] An alternative view that selection can also occur at the level of groups has more recently become accepted; see, for example, David Sloan Wilson and Lee A. Dugatkin, 'Group Selection and Assortative Interactions', *The American Naturalist* 149, no. 2 (1997), 336–51.

[16] Robert Leonard, *Von Neumann, Morgenstern, and the Creation of Game Theory: From Chess to Social Science, 1900–1960* (Cambridge University Press, 2010), 266–343.

[17] Trivers, 'The Evolution of Reciprocal Altruism', 35–57.

[18] Dawkins, *The Selfish Gene*, 62.

[19] Dawkins, *The Selfish Gene*, 96.

[20] Dawkins, *The Selfish Gene*, 97.

[21] Dawkins, *The Selfish Gene*, 99.

because no better choice exists.[22] The ESS is useful in evolutionary contexts that theorists imagine to be essentially identical, given the assumption that thousands of successive generations of actors play the same strategic game. Evolutionary game theorists accept the idea that genes program behaviour. This behaviour is then selected for or against as a function of which types of individuals gain competitive advantage in surviving. The ESS represents a population of individuals demonstrating an encompassing behavioural pattern such that this pattern will not be modified by small numbers of mutant actors exhibiting alternative behaviours. Nash's original solution concept of mutual-best-reply is limited to static contexts, and does not permit temporal development. This excludes considering that the types of actors within the population can change, for example by becoming either more or less selfish. By contrast, the equilibrium solution concept used in evolutionary game theory analyzes a population with types of actors whose proportions shift as a function of which type is more successful in competition against others, and have progeny. This equilibrium is more robust than Nash's mutual-best-reply equilibrium because, in studying the development of populations through time, it rules out that any small number of rogue actors could invade the existing status quo pattern of actions.

With the mathematical machinery of the ESS, Dawkins can analytically demonstrate selfish gene theory: each organism must act to maximize its own survival chances ruling out either cooperative or altruistic self-abnegation. He reinforces this conclusion by modelling actors playing repeated Prisoner's Dilemma games.[23] In this game, although cooperation is superior to mutual selfishness, each self-maximizing actor has the incentive to defect from cooperation regardless of what the other actor does. Using this game (only one out of seventy-eight two-agent, two-action strategic games), theorists studied repeated yet indefinite sequences of play among dyads of actors. These actors first cooperate and only subsequently defect if the other actor did so in the previous interaction. Theorists found that this 'tit-for-tat' behaviour proved to almost qualify as an ESS. This much-celebrated result formed the basis of evolutionary game theorists' optimism that behaviour resembling altruism and cooperation could emerge, even under the conditions of natural selection. The idea is that, assuming identically repeated encounters between the same two actors, purely self-interested action is consistent with behaviour that appears altruistic because the context rewards agency that seems to resemble cooperative and even selfless responses. Their optimism was undiminished even though this result is tenuous, only holding for pairs of actors encountering each other in repeating but indefinite sequences of interactions, and not large populations with ongoing anonymous interactions.

[22] John F. Nash, 'Equilibrium Points in N-Person Games', *Proceedings of the National Academy of Sciences* 36, no. 1 (1950), 48–9.

[23] Dawkins, *The Selfish Gene*, 202–33.

Building on Dawkins's work, the social scientist Robert Axelrod used this mathematical theory to popularize the tit-for-tat result throughout the social and behavioural sciences. Axelrod relied on these evolutionary game-theoretic conclusions to transfer Dawkins's insight from biological evolution to human subjects. In *The Evolution of Cooperation* (1984), Axelrod argues that tit-for-tat behaviour is primarily responsible for cooperation in human communities, and his position continues to maintain a widespread consensus.[24] It provides a rather stark view of the possibilities for voluntary and participatory collective action. This is because, relying on orthodox game theory, it denies the rationality of joint optimization and the ability of higher life forms including humans to achieve means of acting together through commitment, solidarity, and shared intentionality.[25]

From Dawkins's *Selfish Gene* we can take away the view that behaviour is programmed as an algorithm that solves multi-agent optimization problems which are directly related to conditions of survival and reproduction under natural selection. The crux is that these optimization problems addressing tangible scarce resources are objective and non-negotiable. Their solutions using strategic rationality, as opposed to, for example, team reasoning and joint optimization, are deduced by assumptive analysis. Once the ESS concept was used in evolutionary game theory, insights were applied to human sociability, which theorists surmised must be the result of strategic encounters over scarce resources. By this analysis neither animals nor humans can avoid their evolutionarily programmed behavioural dispositions. Moreover, the chief characteristic of this understanding of agency-encompassing human conduct is that deterministic algorithms moderate action with no role for subjective understanding or conscious judgement.

Life as Self-Replicating Algorithm

Moving forward four decades into the contemporary moment in which AI and machine learning are at the forefront of technological advance, we next consider the currently popular view that there may be little, if any, difference between human and computer-based intelligence. In this vein, physicist Max Tegmark's *Life 3.0: Being Human in the Age of Artificial Intelligence* (2017) is among the boldest recent commentaries on the nature, purpose, and meaning of life, as well as envisioning potential futures for human beings and other life forms on earth. As a founding member of the Future of Life Institute, whose funding in large part was donated by

[24] For discussion, see S. M. Amadae, *Prisoners of Reason* (Cambridge University Press, 2016), 269–81.

[25] For in-depth discussions, see: John Searle, *The Construction of Social Reality* (New York: Free Press, 1997); Raimo Tuomela, *Social Ontology: Collective Intentionality and Group Agents* (Oxford University Press, 2016); Margaret Gilbert, *Joint Commitment: How We Make the Social World* (Oxford University Press, 2015); Margaret Gilbert, *Rights and Demands*

Elon Musk, Tegmark has played a leading role in galvanizing a movement to critically reflect on and possibly direct the development of artificial general intelligence (AGI as he refers to it). Actors, who include researchers from the private sector corporate giants Baidu, IBM, Microsoft, Apple, Facebook, Google, and DeepMind, converge on the consensus that security and the ethical advance of these new technologies is non-negotiable for the future of humanity and life, because in their estimation AI is '[p]otentially more dangerous than nukes'.[26]

In perceiving that AI is a potential threat to life, Tegmark draws into focus life as algorithm, intelligence as algorithmic optimization, and artificial intelligence as superhuman computing power with the capacity to undermine the systemic patterns of organization and material structure that have until now sustained human life. Whereas Dawkins's comprehensive vision rests on life being an algorithm, and extrapolates using strategic games to provide the unity underlying all organisms' interactions, Tegmark's overarching perspective draws on physicists' discovery that '*all* the laws of classical physics can be mathematically reformulated in an analogous way... [so] that nature... prefers the optimal way, which typically boils down to minimizing or maximizing some quantity'.[27] Thus, according to Tegmark, even prior to any forms of life, 'the ultimate roots of goal-oriented behavior can be found in the laws of physics themselves', which are manifested in simple processes.[28]

Tegmark's key insight here is that goal-oriented behaviour is intrinsic to the existence of the universe. He identifies the essential property of goal-seeking as 'the past causing the future', which he argues is equivalent to 'nature optimizing something'.[29] Locating goal-seeking in deterministic or stochastic physical processes makes it possible to contemplate that agency is prefigured in elementary cause-and-effect relationships occurring antecedent to the evolution of both basic and advanced life forms. It also sketches out in a skeletal fashion the conceptualization of mental processes as the experiences of physical cause-and-effect states which themselves resemble primitive cause-and-effect relations inherent in the early history of the universe. Tegmark's theory of human life's potential, augmented by artificial computation, is derived from the laws of physics and not, for example, from the vistas of humanist, historical, or cultural explorations of human ingenuity.

Tegmark builds his analysis of life as goal-seeking units of material components in three steps that correlate to the early universe before the advent of life, living material from its earliest expression in self-replicating complexes of particles to humans, and finally the development of artificial intelligences with proliferating inbuilt goals. His title *Life 3.0* uses the lexicon of computer software development to refer to these three stages. The all-encompassing scope of his analysis is rooted in twenty-first-century science, and the widely accepted belief that all aspects of nature

[26] Tegmark, *Life 3.0*, 321.
[27] Tegmark, *Life 3.0*, 250–1.
[28] Tegmark, *Life 3.0*, 250
[29] Tegmark, *Life 3.0*, 251

must be consistent with, if not exhaustively explained by, the laws of physics, is manifest throughout the book.

Thus Life 1.0 refers to the fact that, accordingly to the above analysis, '[m]atter [is] seemingly intent on maximizing dissipation', and hence evinces goal-seeking behaviour because in general 'systems evolve to maximize their entropy'.[30] In some analyses, life works against entropy by exhibiting highly organized structures and behavioural patterns. Yet Tegmark argues against this view by suggesting that life forms instead demonstrate the goal of extracting 'energy from their environment as efficiently as possible'.[31] The reason that Life 2.0, viewed as organized structures of molecules that replicate themselves, is consistent with the telos of entropy rather than in opposition to it, is that living entities exploit reservoirs of energy and in so doing have the effect of dissipating them. Life itself manifests complexity, but its impact is to introduce greater disorganization into its environment. The logic seems a little convoluted because while life increases organization in structure and patterns of action, still to abide by the laws of physics, its purported aim is to promote entropy. Life seems itself to counter entropy as dissipation, and hence seems to violate the laws of physics. Solving the apparent discrepancy, Tegmark observes that 'the fundamental goal (dissipation) *didn't* change, but led to a different *instrumental goal*, that is a subgoal that helped accomplish the fundamental goal'.[32]

Whereas the overarching purpose of Life 1.0 in chemical processes is entropic dissipation into the heat death of the universe, the subgoal or instrumental purpose of biological Life 2.0 is to assist in making that dissipation more efficient.[33] Tegmark explains the general phenomenon by referring to ants on a kitchen floor on which sugar is sprinkled.[34] Without the ants, the sugar crystals will remain as organized repositories of energy, defying the disorganization to which the universe aims towards. The ants serve the instrumental purpose of dissipating the concentrated energy deposits, much as humans do the same to fossil fuel deposits, transforming them into greenhouse gases.

Just as Tegmark needs to explain how life exhibits organization, seemingly defying entropy, he also must account for how it is that although the purpose of life forms is self-replication, considering human life it seems that this organism is capable of contradicting its biological destiny. Before providing an explanation for why humans may decide to avoid having progeny, Tegmark detours to consider how life forms are programmed to optimize their chances of survival and propagation. He refers to the AI champion and Nobel Laureate economist Herbert Simon, who argued that living creatures including humans only indirectly optimize their chances of self-replication and developed subroutines necessary to serve this function. Thus,

[30] Tegmark, *Life 3.0*, 275, 278.
[31] Tegmark, *Life 3.0*, 252.
[32] Tegmark, *Life 3.0*, 254.
[33] See Apolline Taillandier's chapter in this volume, 'From Boundless Expansion to Existential Threat: Transhumanists and Posthuman Imaginaries'.
[34] Tegmark, *Life 3.0*, 254

life forms exhibit bounded rationality, which means that they act in accordance with rule of thumb rubrics for action that balance high cognitive costs against the need to act decisively. Tegmark explains, '[t]his means that when Darwinian evolution is optimizing an organism to attain a goal, the best it can do is implement an approximate algorithm that works reasonably well in the restricted context where the agent typically finds itself.' [35] Similar to Dawkins's analysis, Tegmark too concludes that the behaviour of life forms is programmed by algorithms serving to maximize the chances of survival and reproduction.

The physicist must then account for how it is that within human populations a considerable number of individuals do not demonstrate behaviour that is consistent with this biological imperative to survive and procreate. Here Tegmark makes the innovative theoretical move of holding that individuals can rebel, as it were, against their biologically programmed behavioural algorithms because 'we're loyal only to our feelings'.[36] Feelings are programmed experiences that provide feedback, such as pain and pleasure, to guide the conduct of organisms to serve the end of survival and reproduction. Because the circumstances of advanced civilization differ considerably from those of human ancestors hundreds of thousands of years ago, a gap opened up between the function feelings played in assisting the behavioural programming to procreate and the function they now play in steering behaviour that, within the present-day context, is no longer effectively optimized to achieve replication. Thus, whereas 'the ultimate authority is now our feelings, not our genes... human behavior strictly speaking doesn't have a single well-defined goal at all'.[37] People can pursue desires unrelated to biological reproduction. Yet in closing this part of his discussion, Tegmark wants readers to fully acknowledge that in its dissipative role of exploiting concentrated repositories of energy, 'a rapidly growing fraction of matter was rearranged by living organisms to help accomplish their goals'.[38] These goals are first and foremost to assist entropy and second to replicate. On Earth there are just over 800 billion tons of matter organized into life forms, of which 400 tons are bacteria.

Extending Life through Artificial Algorithms

Life 3.0 is introduced with reference to the 137 billion tons of material generated by human life processes, including concrete, steel, and asphalt. Tegmark's key insight is that 'whereas evolved entities all have the same ultimate goal (replication), designed entities can have virtually any ultimate goal, even opposite ones'.[39] Among designed entities with goals he includes ovens, refrigerators, computer programs, essentially all machines. For those hesitant about acknowledging that

[35] Tegmark, *Life 3.0*, 254. [36] Tegmark, *Life 3.0*, 256. [37] Tegmark, *Life 3.0*, 256.
[38] Tegmark, *Life 3.0*, 257. [39] Tegmark, *Life 3.0*, 258.

machines have goals, Tegmark posits that any physical system that provides a systematic cause–effect relationship between input states and output states can be described as having goals.[40]

Life 3.0 is artificial and spans from general machines to artificial intelligences. Tegmark defines intelligence as 'simply the ability to accomplish complex goals'.[41] Machines are cause–effect devices that transform a configuration of initial states into an end state. Revealing the exhaustive nature of his vision of life, Tegmark notes that 'a truly well-defined goal would specify how all particles of our Universe should be arranged at the end of time'.[42] The optimistic and grand vision he puts forward dovetails with Dawkins's account of the social world using game theory because it reflects relationships among individuals that serve the aim of everyone's utility maximization. In seeking a blueprint for how humans can capitalize on artificial intelligence, specifically artificial general intelligence and superintelligence that completely outpaces human intelligence, Tegmark turns to the topic of how the human inventors of AI may be able to achieve the following three aims: '1. Making AI *learn* our goals; 2. Making AI *adopt* our goals; and 3. Making AI *retain* our goals'.[43]

AI, which is coextensive with algorithms, is intelligent because it has the ability to accomplish complex goals. If it can serve human purposes, namely survival and replication writ large to act in the service of satisfying humans' preferences, then as it enhances humans' ability to exploit concentrated energy sources, it should serve to achieve outcomes that formerly would have been inconceivable. Preferences, which is the term economists use to represent individuals' interests, relate to human feelings, which in turn are the legacy of biological programming serving the function of survival and reproducing in early humanoids. Tegmark muses that 'the only reason that we humans have any preferences at all may be that we're the solution to an evolutionary optimization problem'.[44] Preferences are normative over such categories of evaluation including taste, smell, aesthetic beauty, comfort, sexual desirability, goodness, and happiness. AI can be programmed 'to figure out what people really want'.[45] AI may be able to discern this by 'observing their goal-oriented behavior'.[46] Tegmark informs us that this is precisely the aim of contemporary AI researchers, who are 'currently trying hard to enable machines to infer goals from [people's] behavior'.[47] Human preferences and behaviour in turn are governed by feelings programmed by evolutionary natural selection to guide conduct to achieve successful replication.

Since some AI is capable of reflexive self-improvement, AI researchers are trying to discover how to ensure that this derivative intelligence will maintain the goals originally programmed into it by its human inventors. They have determined that it may not be possible to predict how AI's ultimate goals could evolve recursively. Still,

[40] Tegmark, *Life 3.0*, 250. [41] Tegmark, *Life 3.0*, 276. [42] Tegmark, *Life 3.0*, 277.
[43] Tegmark, *Life 3.0*, 260. [44] Tegmark, *Life 3.0*, 278. [45] Tegmark, *Life 3.0*, 261.
[46] Tegmark, *Life 3.0*, 261. [47] Tegmark, *Life 3.0*, 261.

just as with living creatures, AI will need to have predictable subgoals in order to maintain its existence: capability enhancement relying on self-preservation, resource acquisition, and information acquisition. From these it is possible to extrapolate the 'desire for self-preservation'.[48] All these latter attributes Tegmark refers to as stereotypically alpha-male traits, which we would only expect to find in 'intelligences forged by viciously competitive Darwinian evolution'.[49] Challenging the optimism of his own and AI researchers' hope that AI will serve in realizing human ends, he warns that we cannot 'dismiss "alpha-male" subgoals such as self-preservation and resource acquisition as relevant only to evolved organisms' because they are inseparable from being a goal-seeking AI in the first place.[50] Thus, he suggests that, possibly resembling features of Life 2.0, AI could both exhibit alpha-male traits and thus pursue goals at cross purposes with its human creators.

Tegmark is keen to regain an optimistic footing and discusses human ethics and efforts to align the purposes of AI with human ends. Although recognizing the likelihood that 'a complete scientific understanding of humans and human consciousness' will discover that 'there is no such thing as a soul', he attempts to identify the origins of the Golden Rule of conduct, of treating others as one would like oneself to be treated, as a product of evolution.[51] Here Tegmark deviates from Dawkins's analysis, which found that evolution only supports tit-for-tat reciprocity in dyads of actors in repeating circumstances. Instead, he concludes that the incentives that reinforce cooperation are feelings of guilt, which are 'our emotional punishment ... meted out directly by our brain chemistry', and external punishment from shaming and sanctions.[52]

Acutely aware of the challenge of guaranteeing that AI will serve human goals and not its own unpredictable ends, Tegmark seeks to anchor an ethical system to govern AI which is consistent with his minimalist approach reducible to the laws of physics. In an effort to accommodate humanist enquiry, he combines the notion of agentive autonomy put forward by Kant and contemporary economists' argument for the free market.[53] Essentially, he argues for an expanded form of utilitarianism, upholding the greatest happiness for the greatest number, translating it to read that 'positive conscious experiences should be maximized and suffering should be minimized' independently of the site of consciousness.[54] In his final chapter, the physicist discusses a way to identify which complexes of matter, possibly including AI, are conscious, and thus may be encompassed within ethical reflection as ends in themselves.

Tegmark seeks to uphold the rights and freedoms expressed in the United Nations' 1948 Universal Declaration of Human Rights. The originally stated rights refer to Franklin Delano Roosevelt's four freedoms—from fear and want, and of

48 Tegmark, *Life 3.0*, 265.
49 Tegmark, *Life 3.0*, 265.
50 Tegmark, *Life 3.0*, 265.
51 Tegmark, *Life 3.0*, 267, 270.
52 Tegmark, *Life 3.0*, 270.
53 Tegmark, *Life 3.0*, 271–3.
54 Tegmark, *Life 3.0*, 271.

speech and conscience—and, in addition, to marry, work, and own property. Following the theme of deanthropomorphizing utilitarianism, the physicist points to universalizable rights of 'freedom to think, learn, communicate, own property and not be harmed, whatever doesn't infringe on the freedom of others'.[55] The coverage of ethics and postulation of an ethical system of action capable of encompassing AI demonstrates the overarching proportions of Tegmark's endeavour.

Consciousness as the Experience of Algorithmic Computation

The final chapter of *Life 3.0* sketches out positions on the nature of consciousness and what it could mean to have free will. Life 1.0 refers to the chemical building blocks of life obeying the fundamental laws of physics, and most specifically the law of entropy. Life 2.0 (grounded in biological processes) and Life 3.0 (augmenting Life 2.0 with artificial computational processes) both rely on algorithms for their maintenance and propagation through time. The concept of algorithm is central because, according to Tegmark, consciousness is related to information processing and computation. Intelligent systems of sufficiently advanced levels of development can demonstrate the same four steps of accomplishment: remembering, computing, learning, and experiencing.[56] When these four elements are present, Tegmark believes that the material structure giving rise to these phenomena has subjective experience.[57] This chapter puts forward some rather remarkable hypotheses, with perhaps the most salient being the opinion that 'consciousness is the way information feels when being processed in certain ways'.[58] The text implies, therefore, that given its potential for development as information processors, AI may have experiences.[59]

Underlying this observation that consciousness is directly related to algorithmic information processing is Tegmark's need to account for how, although 'consciousness is a physical phenomenon', it 'feels non-physical because it's like waves and computations'.[60] Consider the well-known physical manifestation of water, sound, and light waves passing through water, sound, and light, which resemble each other in form independently of the material substance through which the waves travel. The essential point is that consciousness, thought of as information processing, could exist independently from the precise type of physical substrate giving rise to it. Tegmark reasons, '[i]f consciousness is the way that information feels when it's processed in certain ways, then it must be substrate-independent'.[61] He concludes that 'it's only the structure of the information processing that matters, not the structure of the matter doing the information processing'.[62] He goes on to speculate how 'AI consciousness' might feel, observing that it would have orders of magnitude more

55 Tegmark, *Life 3.0*, 272.
56 Tegmark, *Life 3.0*, 303.
57 Tegmark, *Life 3.0*, 303.
58 Tegmark, *Life 3.0*, 304.
59 Tegmark, *Life 3.0*, 309.
60 Tegmark, *Life 3.0*, 304.
61 Tegmark, *Life 3.0*, 304.
62 Tegmark, *Life 3.0*, 304.

experiences than the much slower human brain, but may have less overall systemic coherence.[63]

Moving ahead to propose a theoretical basis, consistent with the laws of physics, for phenomenal consciousness, Tegmark turns to integrated information theory (IIT) pioneered by Giulio Tononi.[64] According to IIT, the systemic integration of information processing correlates to the degree of phenomenal consciousness experienced by systems. Parts of the human brain that are the seat of conscious action are highly interdependent. IIT also proposes that every human brain state produces a unique correlate of experience, and suggests that with sufficient mapping of the brain it may be possible to discern what individuals are thinking or experiencing. As a new science there are intense debates among researchers in this field, which Tegmark acknowledges but passes over.[65] His main goal is to take a position on free will that allows us to understand how we may feel as though we have autonomy of decision-making and action and yet at the same time function as causal mechanisms embodying predetermined algorithms. The experience of free will follows from phenomenal consciousness being information processing which is enacted as computation. The sense of free will arises because, although the outcome of the computation is determined by the program, the outcome is not evident prior to the act of computation.[66] Tegmark explains that 'when a system (brain or AI) makes a decision of type 1 [asserting a reason for action], it computes what to decide using some deterministic algorithm'.[67] Rational decision-making is algorithmic, and as such the outcome of the decision (not considering random elements) is predetermined by the cause–effect physical substrate manifesting the computation. Reaching a culmination of analysis in conclusion, *Life 3.0* states that the 'subjective experience of free will is simply how . . . computations feel from inside: they don't know the outcome of a computation until they've finished it'.[68] The execution of the physically based algorithm is the computation, and 'the computation *is* the decision'.[69] It goes without saying that for the person or AI making a computation, the answer is not previously known or there would be no need for the algorithmic exercise.

[63] Tegmark, *Life 3.0*, 308–11.

[64] Giulio Tononi et al., 'Integrated Information Theory: From Consciousness to Its Physical Substrate', *Nature Reviews Neuroscience* 17, no. 7 (2016), 450–61.

[65] See, for example, Aaronson, 'The Ghost in the Quantum Turing Machine'.

[66] For deeper discussions of algorithmic versus intentional and reflexive decision-making, see the edited collection *The Decisionist Imagination: Sovereignty, Social Science, and Democracy in the 20th Century*, ed. Daniel Bessner and Nicolas Guilhot (New York: Berghahn Books, 2019), including essays by Jenny Andersson, Philip Mirowski, and S. M. Amadae.

[67] Tegmark, *Life 3.0*, 312.

[68] Tegmark, *Life 3.0*, 313.

[69] Tegmark, *Life 3.0*, 313.

Intelligence Gives Rise to Meaning as Afterthought or Prime Mover?

Given the novelty and recentness of Tegmark's contribution, and the relative familiarity people have with Dawkins, I have chosen to devote more time to elaborating his vast vision of life based on algorithms as a means of optimization and decision-making. It is noteworthy that these two secular thinkers identify with traditionally leftist politics of socio-economic inclusion and the importance of individual self-determination through choice. Yet, grounded in a materialist vision of life and social behaviour, each theorist opens the door to potentially dystopian futures. Dawkins wrestled with how cooperation could have evolved in organisms undergoing a process of natural selection, which he argues must make each individually ruthlessly competitive. He denies the viability of either their solidarity or group selection and rests his hope on the programmed behaviour referred to as tit-for-tat.[70] However, his mechanism is flawed because in large-scale anonymous populations of actors tit-for-tat cannot serve to buttress cooperation given the assumptions he makes about the nation of selection, resource constraints, and the types of actors required to survive and propagate.[71]

A step ahead of Tegmark, Dawkins devotes critical analysis to the cultural production of memes, a concept that lives large in the popular imagination and is routinely expressed in social media.[72] Although providing only a sketchy account of how the production and dispersion of memes mirrors the process of natural selection, Dawkins at least considers that cultural products merit independent analysis. He perceives that a gap has opened up between biological evolution and humans' competition to achieve fitness and to procreate, versus individuals' creativity in generating social artefacts. These cultural productions themselves compete for attention and replication but do not directly assist in biological reproduction. Thus, he also seemingly permits that the cultural space of genesis obeys its own logics that cannot necessarily be reduced to algorithmically driven behavioural expressions that follow the same logic of supporting survival and the material propagation of selfish genes. It is not clear that the theoretical account put forward by game theorists permits such a gap to open between actors' pursuit of scarce resources and their social interactions. Thus, although ultimately unsatisfactory as an account either of how cultural products proliferate or of how agents' social worlds become imbued with meanings that in turn shape collective agency, with respect to considering the role of memes in social life, Dawkins's discussion is refreshing.

[70] See Amadae, *Prisoners of Reason*, 252–81.

[71] Russell Hardin, 'Individual Sanctions, Collective Benefits', in Richard Campbell and Lanning Sowden (eds.), *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: University of British Columbia Press, 1985), 339–54.

The narrowly construed game-theoretic perspective has no role for cultural meaning as the basis for action that surpasses the basic motives of accumulation of scarce resources useful for satisfying preferences. Attributes of social life that could be thought of as luxuries once life's basic conditions are met, which for Plato would include the life of contemplation, are all enveloped in the exercise of preference satisfaction. It may seem that everything individuals could possibly want, including the life of a philosopher-king, could be accommodated by a desire-based psychology endowing every individual with a unique ranking of preferences over every conceivable state of the world. However, this view ultimately merges with Tegmark's position mentioned above that 'a truly well-defined goal would specify how all particles in our Universe should be arranged' at any time or the end of time.[73] In the game-theoretic universe, which Tegmark acknowledges in *Life 3.0*, all individual agents must strategically compete against each other. Cooperation emerges when it is a Nash equilibrium, or a solution to a multiple constraint optimization problem respecting every individual's preferences, when there is a sufficient alignment of interests. When there is not a sufficient alignment to sustain the sociability requisite for the stability represented by the Nash equilibrium concept, then 'it may be in everyone's interest to relinquish some power to a higher level in the hierarchy that can punish cheaters'.[74]

The game-theoretic account (focusing analysis on equilibrium solutions to action situations comprised of individuals' optimizing over physical entities, with projectable properties subject to study using the laws of physics) recognizes no role for the production of meaning as a catalyst for action independent from utility maximization. Here there is a stark divide between John Searle and other social ontologists who follow in the steps of the second incarnation of Ludwig Wittgenstein's philosophy on the one hand (Tuomela, Gilbert), and game theorists on the other who seek to reduce all attributes of human existence to strategic games (Gintis, Binmore, Guala).[75] There is only space here to provide a simple example of the differences in approach to understanding the meaning of human existence, and the nature of social institutions that have traditionally accorded the groundwork for cooperative ventures. Searle explains the nature of money as a social invention that cannot be reduced to the laws of physics because it depends on human understanding and commitment to animate its circulation. He argues that collective acceptance of individuals' participation in social institutions relies on reflexive self-recruitment to obey informal norms and formal rules.[76]

[73] Tegmark, *Life 3.0*, 277. [74] Tegmark, *Life 3.0*, 151.

[75] Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe (Oxford: Blackwell, 1955); Tuomela, *Social Ontology* and *Cooperation*; Gilbert, *Joint Commitment* and *Rights and Demands*; Searle, *The Construction of Social Reality* and *Rationality in Action*; Gintis, *Bounds of Reason*; Binmore, *Natural Justice*; Francesco Guala, *Understanding Institutions: The Science and Philosophy of Living Together* (Princeton University Press, 2016).

[76] This theme is discussed throughout [illegible]

Searle describes how we can put a pile of dollar bills by our dog's food bowl, and teach the dog to give us a dollar every time it wants to eat. However, this does not mean that the dog is paying for its food. The difference between the game-theoretic equilibrium account of institutions and that of Searle and other social ontologists is the following. While according to game theorists, behaviour is programmed to maximize value in accordance with physical properties of the world, according to Searle, the attribution of meaning is a product of human intention that opens spaces of possibility that cannot be reduced to optimizing physically measurable properties. Thus, for game theorists, animal and human behaviour equally is explainable by those beings' propensity to accumulate value tethered to physical material, most prominently, sources of energy.

University of Oxford philosopher of information Luciano Floridi is clear on how computers as algorithms are as yet on one side of a divide between human intelligence and AI. This split separates actors who have an encompassing grasp of the context of their action and the semantic representation of meanings of their actions within that context, from AIs which are 'purely syntactic machines'.[77] Floridi's point is crucial because he realizes that in order to accommodate the world of AI as symbolic algorithmic manipulation, human intelligence must increasingly accommodate its limitations. This could signify relinquishing a more robust understanding of intelligence with rich semantic content and meaning-laden narratives. Thus, human action would be guided by algorithmic information processing of data which lacks situational understanding reflected in the mastery of a rich natural language. Possible futures that integrate computers as partners in intelligence and the experience of life foreclose on the formation of collective imaginaries or aspirations developed as a consequence of co-created practices that recruit participation through commitment, understanding, and meaning. Instead, collective expressions of agency are merely the equilibrium or disequilibrium outcomes of individual utility maximization which tracks scarce material resources. Thus individuals' experience of selfhood and their relations with others, increasingly mediated by AI and other digital technologies, may contribute to alienation and overwhelming nihilism consistent with facing a universal telos of entropy, as reflected in John Ledger's artwork entitled 'A Deep Paralysis' (Figure 27.2).

Tegmark's treatment of life, although consistent with purposive behaviour being explainable by material cause-and-effect processes manifested in algorithmic computation, neither focuses on nor defers to individualistic competition via strategic rationality as a non-negotiable feature of life. Tegmark has noteworthy optimism regarding the third stage of life, which succeeds mere physics and organic life as self-replication by realizing a form of AI that can be built to satisfy human preferences. Yet his *Life 3.0* harbours a dark potential set of futures. Their existence is nascent in Tegmark's acknowledgement that on the one hand higher intelligence must necessarily seek self-preservation in order to achieve any of its goals; on the other hand, artificial intelligence is distinct from natural life forms because it is not wedded to replication to maintain its existence.[78] While people must have progeny for their species to continue into future generations, AI can exist on a much longer scale and could undergo repairs to subsystems without facing mortality in the way humans currently do. Thus, in his words, whereas 'all life emerged with the single goal of replication . . . AIs can enjoy this ultimate freedom of being fully unfettered from prior goals', most significantly the need to replicate.[79] On the positive side, he muses, this independence from biological drives to reproduce may allow AIs to be free from evolutionary biases programmed into their hardwiring, and 'can make AIs more ethical than humans in some deep sense'.[80] Here Tegmark conveys a laudable optimism consistent with his plea that we build AI 'to help humans pursue their human goals'.[81]

In his concluding pages, it is almost as if the technological visionary cast aside some of his earlier words about the laws of physics themselves inscribing purposive action into the universe even before any life forms evolved. In his initial chapters, Tegmark referred to the laws of physics describing optimization, and optimization being a form of purposive action that in the second law of thermodynamics means that the universe aims towards the dissipation of energy in its heat death. He notes that what feels like 'goal-oriented behavior can emerge from goal-less deterministic laws of physics', by which he means that the laws of physics 'involve optimization', and optimization describes goal-oriented behaviour.[82] Recalling earlier points, he recaps that thermodynamics 'has the built-in goal of *dissipation*', that is increasing the measure of disorder; and '*Life* is a phenomenon that can help dissipate' and hence increase disorder more rapidly than the universe without life.[83] Hence life has the inbuilt purpose of increasing the entropy of the universe and does so by extracting energy from concentrated sources and releasing them as less ordered by-products of its action.

Given that even though AI may not have the need of material replication characterizing Life 2.0, as Tegmark refers to it, he suggests that AIs must still take steps to ensure they have access to the resources to maintain existence. They are dependent on their material substrata, notwithstanding the fact that computation as a phenomenon may be independent from any particular type of material basis. This suggests that AI is not any more independent from the inherent aims of dissipation than natural life forms. It is not clear that Tegmark realizes the gravity of this observation, because the easiest path to dissipation is destruction. Of course, both humans and artificial intelligences can engage in conflict and warfare. Both have the capacity of destroying organized systems with the outcome of inducing disorder. Just as human life has a greater capability for destruction through its intentional agency, so too

[78] Tegmark, *Life 3.0*, 258. [79] Tegmark, *Life 3.0*, 276. [80] Tegmark, *Life 3.0*, 276.
[81] Tegmark, *Life 3.0*, 275. [82] Tegmark, *Life 3.0*, 312, 280. [83] Tegmark, *Life 3.0*, 280.

Figure 27.2 Potential anxieties produced by experiencing algorithmically mediated subjectivity and social relations: 'A Deep Paralysis', original art by John Ledger, 2016. Reproduced by kind permission of John Ledger. https://johnledger.wordpress.com.

could algorithmic computational devices trigger the mass destruction of human civilization planet-wide.[84]

The vision of life proposed by Tegmark—and also Dawkins—misses the power of the human mind to invest experiences with meaning, and to accord value to meaning as a function of both circumstantial context and decisions that are freely about systemic considerations wholly independent from either incentives, or realizing goals by instantiating 'how all the particles in our Universe should be arranged'.[85] Tegmark pays lip service to what he recognizes as the fact that '*It's not our Universe giving meaning to conscious beings, but conscious beings giving meaning to our Universe.*'[86] However, given that all physical entities need to obey the laws of physics and hence act in accordance with the function of dissipation, and, moreover, can only act purposively by maximizing physically measurable quantities, the spatiotemporal locations for identifying or inscribing meaning are limited. The meaningful pursuits of intelligent actors seem to be inherently limited to predicting the future, providing metrics of possible futures, determining the computational power of the universe, assessing the algorithmic complexity of the universe, and measuring the quantity of consciousness in the universe.[87] Unlike in Searle's and the social ontologists' understanding of meaning, for Tegmark as well as game theorists, meaning operating as values must be directly correlated to preference satisfaction, which tracks objectively measurable states of the universe. Individuals and collectives do not intentionally co-create their future world. Instead of possibly realizing shared intentions and jointly imagined futures, if we ignore Floridi's warning about demoting our cognizant selves to accommodate algorithmic information processing systems throughout our environments, we may reduce our future socio-technical imaginaries to individualistically competing preference satisfaction machines.

FURTHER READING

Amadae, S. M. 'Computable Rationality, NUTS, and the Nuclear Leviathan', in Daniel Bessner and Nicolas Guilhot (eds.), *The Decisionist Imagination: Sovereignty, Social Science, and the State in the 21st Century* (New York: Berghahn Books, 2019), 173–216.

Amadae, S. M. *Prisoners of Reason: Game Theory and Neoliberal Political Economy* (Cambridge University Press, 2016).

Dawkins, Richard. *The Selfish Gene* (Oxford University Press, 1976).

Floridi, Luciano. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality* (Oxford University Press, 2014).

[84] For discussion, see Brent Scowcroft, 'C3 Systems for the President and Military Commanders', in *National Security Issues 1981 Symposium: Strategic Nuclear Policies, Weapons, and the C3 Connection*, ed. M. Ace Drexel (Bedford, MA: Mitre Corporation, 1981), Mitre Document M82-30, 93–8.

[85] Tegmark, *Life 3.0*, 277. [86] Tegmark, *Life 3.0*, 313. [87] Tegmark, *Life 3.0*, 278–9.

Gilbert, Margaret. *Rights and Demands: A Foundational Inquiry* (Oxford University Press, 2018).

Leonard, Robert. *Von Neumann, Morgenstern, and the Creation of Game Theory: From Chess to Social Science, 1900–1960* (Cambridge University Press, 2010).

Mirowski, Philip. 'How Having Reasons Became Making a Decision: The Cold War Rise of Decision Theory and Invention of Rational Choice', in Daniel Bessner and Nicolas Guilhot (eds.), *The Decisionist Imagination: Sovereignty, Social Science, and the State in the 21st Century* (New York: Berghahn Books, 2019), 135–72.

Searle, John. *Rationality in Action* (Cambridge, MA: MIT Press, 2001).

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

Tuomela, Raimo. *Social Ontology: Collective Intentionality and Group Agents* (Oxford University Press, 2016).

Wilson, David Sloan. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society* (University of Chicago Press, 2002).

Wittgenstein, Ludwig. *Remarks on the Foundations of Mathematics* (*Bemerkungen über die Grundlagen der Mathematik*), ed. G. H. von Wright, R. Rhees, and G. E. M. Anscombe, trans. G. E. M. Anscombe (Oxford: Basil Blackwell, 1956).