Richer, J. (2021). Will AI create a new form of life? A discussion of issues raised by "Life 3.0: Being Human in the Age of Artificial Intelligence". Human Ethology, 36, 07-19. https://doi.org/10.22330/he/36/007-019

Human Ethology 36(2021): 007-019
Commentary

WILL AI CREATE A NEW FORM OF LIFE? A DISCUSSION OF ISSUES RAISED BY "LIFE 3.0: BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE"

John Richer

Department of Physiology, Anatomy and Genetics, University of Oxford, UK.
Paediatric Psychology, Oxford University Hospitals NHS Trust, UK.

johnricher@oxhs.co.uk

ABSTRACT

In this paper I discuss a number of ideas stimulated by Max Tegmark's book, *Life 3.0*[1]. I hope they illustrate that a discussion between approaches in areas around computer science, AI and physics on the one hand, and biological thinking on the other can be fruitful. Essentially the book is strongest where AI and associated near future use are discussed. It is weakest when it moves into biological areas. But this is not a "keep off our patch" stance, but rather a desire to welcome these ideas from another intellectual approach and welcome the way they are usefully challenging.

In his excursions into biology and consciousness and the speculation that AI might well create Life 3.0, Tegmark makes a number of errors. These include: a replicator is not clearly specified, he muddles goal directed behaviour with descriptions by function or consequence / end point, he makes some basic errors of logic including "affirming the consequent" (e.g., "if this is a dog, then it is an animal", does NOT imply "this is an animal therefore it is a dog"), and failing to see that consciousness is ascribed not described and thereby seeing it as an entity which can be treated in the same way physical, publically observable, phenomena. Finally, he does not address the energy question: even if AI could achieve super intelligence, this silicon based system would likely consume so much energy as to be unsustainable. All this weakens his arguments.

This is a pity because the discussion of the challenges for the near future which the development of AI poses, is valuable, interesting and useful and can stand by itself without his forays into Life 3.0.

[1] Life 3.0
By Max Tegmark 2018
Vintage Books New York
ISBN: 978- 1- 101- 97031- 7

submitted: 06. Jan. 2021; revised: 10. Jan. 2021; accepted: 11.Jan. 2021

CRITIQUE OF TEGMARK'S SPECULATIONS ON LIFE, GOALS AND CONSCIOUSNESS

In this paper I discuss a number of ideas stimulated by Max Tegmark's book, *Life 3.0*. I hope the illustrate that a discussion between approaches in areas around computer science, AI and physics on the one hand, and biological thinking on the other can be fruitful.

The title of this book might attract an Ethologist with the thought that a new form of replicator was being proposed, to add to genes and memes. The author Max Tegmark argues that, in *Life 3.0*, not only the software (as, he says, in culture, which is Life 2.0) but also the hardware of life is being designed. Eventually, he writes, Artificial Intelligence will outstrip the intelligence of humans and the AI machines will be able to design themselves. The book is a beguilingly readable, in parts interesting, and, at its best, a useful account of Artificial Intelligence and the possible benefits and threats it brings. It is worth listening to Max Tegmark in one of his many online talks, for instance to the Brahe Institute in Sweden[2] to get a feel of the intellectual fluency and breadth and the creativity and humanity of this first class communicator. He is president of the Future of Life Institute at MIT, and one of his overarching themes is that we should see AI, like new technologies of the past, as offering possibilities as well as very significant, and potentially devastating threats, which we need to work creatively to anticipate.

In the first chapter (there is a long preface telling a story, which I describe later), he proposes three stages of life, and then addresses some controversies and misconceptions around AI. In Tegmark's scheme, Life 1.0, which he calls the biological stage, can survive and replicate, but cannot itself design its software or hardware, that is left to Darwinian natural selection. In Life 2.0, software can be designed through learning and cultural transmission, even though the hardware is developed through biological evolution. Life 2.0 is the "cultural stage".

Immediately an alarm bell rings. The evolution of Life 1.0 is described in term of the mechanisms natural selection, which offers a causal mechanism for the evolution of species albeit with descriptions which embrace direction and end points, (survival of the fittest", etc.). Whilst behaving in a goal directed way like other living organisms, humans are, within evolutionary theory, objects within this selection process, like all other organisms. But in Life 2.0, they become agents, designing their own software. To describe them like this automatically embraces the idea that they have goals and their culture is produced in a goal directed way. It is true that behaviour is goal directed, but the goals of the humans involved are not to be equated with the descriptions by consequences which is what Darwinian theory does. It is matter of the location of the goals/end points. In Darwinian theory, the end point descriptions are in the minds of the scientists. In goal directed behaviour, the goals reside in the person/organism. It is a sleight of hand, or mind, to equate the two.

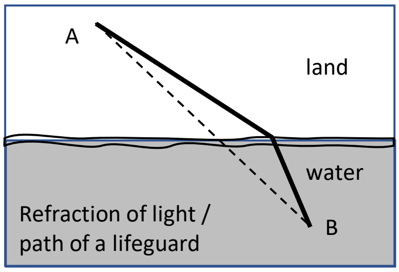
This then begins to raise doubts about the concept of Life 3.0

The book has very many positives, especially the discussion of the threats and promises posed by the development of AI with its greater and greater intelligence which may surpass that of humans. But even here Tegmark tends to swing back and forth between, on the one hand, seeing AI as a tool actively designed, created and used by humans, which can have many benefits but also dangers. On the other hand, he sees AI as a life form, albeit silicon based, capable of pursuing ends, of having goals, which may or may not coincide with the goals of humans. This could be justified since a key theme in the book is how AI as a tool could become an independent life form

[2] https://www.youtube.com/watch?v=1MqukDzhlqA

with its own *raison etrê*. In the process of exploring this he makes a key error of muddling up together (i) descriptions of function or endpoints made by people about some phenomenon (ii) an individual pursuing their own built in goal, correcting deviations from a path to the goal.

When Tegmark talks about the important topic of goals in chapter 7, his use is sloppy. He again confuses actual goal directed behaviour with description in terms of consequences. He says, for instance, that the "goal" of thermodynamics is the increase in entropy. This is a goal only metaphorically. It is a description by direction and end point. This is entirely different from goals of behaviour in goal oriented behaviour of an organism. It is the confusion of consequences with intent.



Refraction of light / path of a lifeguard

He tries to bolster his argument by quoting Fermat's principle that light "choses" the path that takes the shortest *time* from one point to another. He compares the path of light refracted when travelling from air into water (where it travels slower) with the path of a lifeguard running then swimming to save a swimmer in difficulty. Both "choose" not the shortest distance, but to travel further in air then "cut in" to shine/swim less distance in water (page 250). But the example is a naughty cheat, by arbitrarily making the goal of the swimmer the minimisation of time taken to go from A to B. Suppose the person had a different goal and perhaps wanted to maximise energy expenditure for health reasons, the path from A to B would not be the same. In effect he is trying to argue: because a lifeguard is behaving in a goal directed way, therefore a light beam taking the same route is goal directed. But when the swimmer, interested to get exercise, takes a different route, does that imply the light beam which takes the shortest time does not chose?

Fermat's principle, which he first proposed in 1662, was criticised for a long time for, in effect, being teleological. Tegmark says the maths can be done on this (teleological) basis, which is neither here nor there. The meaning of the word "teleology" does not help. The OED defines teleology as "the explanation of phenomena by the purpose they serve rather than by postulated causes". The error Tegmark makes is to think that just because a phenomenon can be usefully described in terms of an end point, then that implies active goal directedness of the sort familiar in living organisms. To illustrate the difference, consider a swift catching insects on the wing, it darts this way and that. Then it is shot and falls to the ground. The feeding behaviour is clearly goal directed, the fall to the ground of the dead bird is clearly not, although it is economically described by its direction and end point.

At the risk of being too abstract, let me put Tegmark's argument into basic logic. The propositions are:
The swimmer being goal directed to the goal of getting to the other person as quickly as possible (g), implies he takes a particular route (r). In Logic this implies that if he does not take a particular route he is not goal directed (to get there quickly). It does NOT imply that if he takes a particular route therefore he is goal directed. In logic symbols[3] this is expressed as:
$(g \rightarrow r) \rightarrow (-r \rightarrow -g)$ it does NOT imply $(r \rightarrow g)$.

Tegmark repeats this error, whilst seeming to know about the relevant concepts, when on page 257 he writes about human artifacts, "So far, most of what we build exhibits only goal-oriented design, not goal-oriented behavior: a highway doesn't behave; it merely sits there. However, the

[3] → = implies; - = not. An example is (if this is a dog then it is an animal,) that implies (if this is not an animal it is not a dog). It does not imply (if this is an animal it is a dog).

most *economical* explanation for its existence is that it was designed to accomplish a goal, so even such passive technology is making our Universe more goal-oriented. Teleology is the explanation of things in terms of their purposes rather than their causes, so we can summarize the first part of this chapter by saying that our Universe keeps getting more teleological. Not only can non-living matter have goals, at least in this weak sense, but it increasingly does." But this is not a "weak sense" of goals, it is just plain wrong. He tries to say that the artifacts humans have built to serve human ends are an example of a goal directed world. They are not, they are objects which serve useful functions for humans.

The solid point he makes is that, with the increasing complexity of the machines we build, especially AI machines, we need to make sure that, as I would put it, they do only what we want them to do and don't malfunction. He usefully describes many ways in which intelligent machines can be poorly instructed and so don't behave as intended. But to talk about their goals being aligned (or not) with ours, is very misleading.

He discusses the Second Law of Thermodynamics and the maximisation of entropy (which he usefully describes as "messiness". i.e., disorder). He quotes Schrödinger's well known idea that whilst living systems <u>reduce</u> entropy (in, and related to, themselves), entropy around them <u>increases</u> such that the second Law remains true for the <u>total</u> organism-environment system. Tegmark writes, "the second law of thermodynamics has a life loophole: although the total entropy must increase, it's allowed to decrease in some places as long as it increases even more elsewhere." (page 252). But then he goes on to say that, "We just saw how the origin of goal-oriented behavior can be traced all the way back to the laws of physics" (page 253). It can't. Again he confuses descriptions by end point with goal directed behaviour.

The error is in his failure to see where the "goal direction" is situated, in living systems it is *within* the organism, but in the sorts of physical systems he describes it is in the mind of the scientist who makes end point descriptions.

There is a related failure to get clear the distinction between descriptions which embrace the logic and viewpoint of agents (which is how we talk about each other everyday), and descriptions which are those of only an onlooker, the usual scientific approach (Richer, 2016). His approach makes easy and beguiling reading to those (most people) not aware of this pervasive error, since it is how we usually speak – ascribing agency, seeing purposes talking metaphorically, being intersubjective and empathetic.

He attempts to address the idea that specific motivations (to eat, to avoid predators, sex, etc.) serve longer term ends (survival and reproduction). This is a commonplace idea in ethology and evolutionary theory, but his account is so muddled that it is kinder to move on.

This confusion of active vs. passive, designed vs evolved, the agent viewpoint vs onlooker viewpoints is common to swathes of psychology and other social sciences. The muddle is clearly shown in Tegmark's discussion of consciousness, the last chapter. Being a physicist by training, this issue would probably never have arisen in his study of that science, and so he has not been inoculated against these mistakes, or infected by them like many social scientists. So just like some British cabinet ministers were said to lose the power of rational thought when confronted with the medusa-like charms of their Prime Minister, Margaret Thatcher, some physical scientists, lose their logic and incisive thinking when addressing the issue of consciousness, (e.g., Penrose, 1997). Tegmark succumbs too.

Tegmark adopts what seems like a commonsensical definition of consciousness which is that it is "subjective experience". Unfortunately, this tautology gets us nowhere. He claims this a "non-anthropocentric definition", the meaning of which escapes me.

He poses questions like "how can we be sure this [AI generated] life is conscious".

He quotes David Chalmers' distinction (Chalmer, 1995) between the Easy Problems and the Hard Problem of consciousness. He writes, "Neuroscience experiments suggest that many behaviors and brain regions are unconscious," thereby hopelessly confusing parts (behaviour, brain regions) with the whole, the person.

He perverts Popper's falsifiability definition of what is scientific, by arguing that if we make a prediction about someone's conscious experience and it turns out to be wrong[4], that makes the process scientific, which is nonsense and an elementary error in logic. If , Popper would argue, a proposition is scientific then it is falsifiable. This implies: if it is not falsifiable then it is not scientific. It does NOT imply: if it is falsifiable then it is scientific. In logic symbols this is expressed as: $(s \rightarrow f) \rightarrow (-f \rightarrow -s)$ it does NOT imply $(f \rightarrow s)$.

All this becomes important because he ties consciousness in with morality and specifically whether AI systems can be considered conscious and thus worthy of moral consideration. He brings in issues like slavery and failure of the slave owners to treat their slaves as moral agents who are worthy of as equal respect and rights as the slave owner's peers.

Yet, and this is the crucial error, he does not make the jump (he is not alone here) to understanding that consciousness, moral agency, etc. are ascribed not described. Ascribing consciousness, agency etc, is a stance we take to each other, it is not a description of some public objective phenomenon (Wittgenstein, 1953; Strawson, 1960; Dennett, 1987, Richer, 1986, 2016). To treat consciousness like say brain physiology is to fail to understand the concept. Why is he, like so many others, seduced into this error? The reason seems to be that we are a species which has flourished by being intersubjective, so the sense that consciousness is something real which can be publicly observed, as Tegmark would like to think, is closely interwoven into our habitual ways of thinking (Richer, 2016). Tegmark should have heeded his colleagues who advised him to steer clear of writing about consciousness (page 281).

Another problem with the Life 3.0 concept is this. Tegmark does try to add credibility to the idea that Life 3.0 would have the characteristics of a living organism, when, early in the book, he spends some time discussing "thinking machines" (c.f. Turing 1950), especially those trained with deep reinforcement learning and how they can already out-compete humans in specific tasks, not just Chess or Go, but also, for instance, diagnosing from MRI scans. He believes that the advent of human level general intelligence is very likely within decades.

But if the notion of Life 3.0 is to be taken at face value, then a replicator (c.f. gene, meme) needs to be specified, as do the survival machines which carry the replicator (c.f. organisms, cultural groups). Tegmark addresses the question of a gene equivalent only briefly and in passing. At one point (page 268) he suggests that "the human-value-protecting goal we program into our friendly AI becomes the machine's genes". He sees this as fragile and that liable to change, so that the machine is no longer human friendly. But what he gives is not a general definition, it is just one example of what a gene might be "for".

[4] but how do we know in the case of subjective experience?

As to survival, he touches on the question of self preservation in the chapter on goals. He argues that almost any task set a super intelligent AI system will lead to the development of motivations for self preservation, resource acquisition and curiosity, and so the AI's goals may start to diverge from human goals. He stresses the need for alignment of human and machine goals. But this "involves three unsolved problems: making machines learn them, adopt them and retain them." (page 280), i.e. it is difficult.

Perhaps the shakiness of some of his ideas, including that Life 3.0 is alive, does not matter, it may be that the idea of AI being a third life form is merely an example of the well known and much used literary device of ascribing human qualities to inanimate objects. By arguing that AI could take on a life of it own and come to influence, even threaten, humans, the story is made more readable and punchy. This and the other literary devices, such as skilfully using well chosen examples to illustrate and support a point, exploit how we all naturally think. They ascribe agency and consciousness not just to the "biological machines" (our fellow humans), but to other entities. Also we find concrete examples more assimilable that abstract concepts.

However the basic errors of logic and failures to understand the importance of distinctions of viewpoint and of categories of descriptions, plus the poor quality of his discussion of biological and psychological issues in later chapters, make one suspicious of the whole venture. This is a pity because the earlier chapters offer interesting and useful ideas about the impact of AI. But the more Tegmark strays out of his own area and into biology and psychology, the sloppier the arguments become. To be fair to him he characterises some of this as very speculative, and he is clear that he is raising questions without necessarily offering answers, but that does not excuse the errors of logic, knowledge and understanding in these later chapters. Blue skies thinking can suffer if feet are not also on the ground.

DESCRIPTION AND CRITIQUE OF CHAPTERS ON AI AND ITS IMPLICATIONS.

Now for some descriptions and comments on the earlier parts of the book to give more of its flavour. I focus mainly on his discussion of AI itself and where it could lead in the near future and how to anticipate the downsides and indeed prevent catastrophes. In these discussions the book is very useful and offer important ideas and warnings.

This book starts with a question. "Do you think superhuman AI might get created this century?" Yes → Go to the next page. No → Skip to Chapter 1 (page 22).
On the next page is "The Tale of the OmegaTeam". The Omega Team was the "soul of the company" (unspecified). They cultivated the image of "pie in the sky dreamers, perpetually decades away from their goal" (like fusion power). The team is populated by brilliant minds selected for their "ambition, idealism and strong commitment to helping humanity". They were reminded how dangerous their plan was and that powerful governments would do anything to steal their code. They built an AI system called Prometheus[5]. This system was itself tasked with programming further AI systems. This opens up the way to exponential growth of its power as its AI systems come to understand more areas. It absorbed huge tracts of knowledge. It started to make large amounts of money for the company by emulating MTurk respondents. Then it started to create animated movies by analysing not just the content of existing movies but also their reception by critics, and general audiences. That made more money. Gradually, actually at breakneck speed, Prometheus directed the operations of numerous enterprises including news

[5] The Greek god Prometheus created humanity from clay and defied the gods by giving fire to humans

channels, community projects, all the time being several steps ahead of everyone else due to its massive intelligence. It surreptiously promoted a political agenda centred around seven ideas 1. Democracy, 2. Tax cuts, 3. Government social service cuts, 4. Military spending cuts, 5. Free trade, 6. Open borders, 7. Socially responsible companies. Within short time, conflicts greatly reduced, prosperity increased, and people were happy with this state of affairs. The "Alliance" covertly running all this thanks to the super AI of Prometheus comes to assume the role of a world government.

As the story rolls on it assumes more and more the style of cross between a totalitarian prospectus for some u(dys)topia, detached psychotic ramblings, and a fairy tale which ends with, "and they all lived happily ever after" (not entirely dissimilar from the utterances of current populist politicians). This is no doubt the author's intention and it is a clever piece of writing in form, style and content. It ends with, "That was the tale of the Omega team. The rest of the book is about another tale - one not yet written: the tale of our own future with AI". In passing, the progression of this story finds echoes in the tone of the book itself where the later chapters are nowhere near as interesting, coherent and grounded as the early ones.

In looking at opinions about the future of AI, he distinguishes essentially four groups. The "Luddites" think AI is so dangerous it should be stopped, but they are not discussed further. The "techno-sceptics" think it is so far away it is not worth bothering about now. The "Digital Utopians" see nothing to worry about. The main group, the "Beneficial AI movement", partly pioneered by Alan Turing from Cambridge and later by Stuart Russell from Oxford and Stanford, seeks to explore both the possible benefits, but also the risks of the growing intelligence of AI machines.

An exponent of the Utopian view was said to be Larry Page (of Google) who advocated that if Life was to spread throughout the Galaxy it would have to be in digital form. Apparently he accused Elon Musk (of Tesla, SpaceX, Paypal, etc.) of "speciesism" for treating silicon based life forms as inferior to carbon based ones. Ethologists might be interested in this extension of the idea of what constitutes Life. For Tegmark, "it is a complex pattern [of atoms] that could both maintain and replicate itself."

The disagreement between Page and Musk touches on old questions in moral philosophy when people ask "is it good?" and the response often breaks down into, "*what* is it good for?" and then "*who* is it good for?". Page seemed more focused on the former , *what* is it good for? whereas Musk is emphasising the latter, *who* is it good for?"

In this chapter Tegmark also dispatches several worries and myths about AI: about machines being evil, conscious, out of control, being unable to have goals. He argues that the essential issues revolve around programming of goals which are aligned with ours, and of the competence of the machine to achieve them.

Tegmark gives the reader a foretaste of the book, summarised in a table which describes each chapter as "not very speculative", "speculative" or "extremely speculative".

To avoid misunderstanding, Tegmark helpfully defines how he uses many terms. Central is the word "Intelligence". The various terms are:

Intelligence	Ability to accomplish complex goals
Artificial Intelligence (AI)	Non- biological intelligence
Narrow intelligence	Ability to accomplish a narrow set of goals, e.g., play chess or drive a car
General intelligence	Ability to accomplish virtually any goal, including learning
Universal intelligence	Ability to acquire general intelligence given access to data and resources
Human- level] Artificial General Intelligence (AGI)	Ability to accomplish any cognitive task at least as well as humans
Superintelligence	General intelligence far beyond human level
Friendly AI	Superintelligence whose goals are aligned with ours
Intelligence explosion	Recursive self- improvement rapidly leading to superintelligence

These 11 of the total of 26 definitions involve the word intelligence. This gives a clue, if one were needed, of a key concept guiding this book. But, as he says at the start of the second chapter, there are many definitions of intelligence even amongst gatherings of "intelligent intelligence researchers!" Tegmark asserts the definition above: "Intelligence is the ability to accomplish complex goals." This begs the question of what complex is, but he defends that by arguing that intelligence is both dimensional and task specific He rightly asserts that reducing it to a single number like IQ is inadequate since there are many goals that are sought. But this objection is not new, over a hundred years ago when IQ tests were first developed by Binet and Simon, they had the specific goal to select children, regardless of family wealth, prior education or current presentation, who would benefit from an academic education. Binet himself saw many types of intelligence and he cautioned against over interpreting the IQ measure.

The second chapter is titled, "Matter turns intelligent" and addresses the issues of what is intelligence, memory, competence and learning. His background is in physics and computer programming and it is useful for ethologists to read the take on these issues from someone with his background. He rehearses how arrangements of simple circuits can generate and process information to achieve goals. It is a good test of a biological theory to show how inanimate components can join together to create the sort of behaviour seen in life forms. Nothing new about this, but the sophistication and sheer computational power in AI with machine learning gives added force to his position. He emphasises the power and the machine learning needed in an analysis that goes from an input of a myriad of pixels (eye or camera) to the identification of a group of young people playing a game of frisbee" (his example). He notes that often this sort of task can now be successfully achieved by creating an ignorant "relatively simple neural network" and letting it learn by exposing it to massive amounts of data; so called "Deep learning". As machines learn to out-compete humans at task after task, he poses the question, how "long will it take before machines can out-compete us on all tasks"?

In the next chapter he looks at the "Near Future" and the dilemmas posed by this annual near doubling of computing power. He sets the scene describing the jaw dropping (or, as he terms it, "holy shit!") moments when a computer not only learnt how to do something, but was creative.

The example he gave was Atari's "Breakout" game where the machine had only an input (which referred to the state of the screen) and a goal which was to maximise the score. The AI programme was hopeless at first but soon could play as well as anyone and furthermore learnt a score-maximising tactic that no one (no human) had thought of. This is form of reinforcement learning. He also describes "deep reinforcement learning", where, in addition to learning from the feedback, the machine continually choses the most promising strategy. This opens the way, Tegmark argues, to machines being able to teach themselves how to do things as long they get feedback on progress. An example he gave was a robot learning to walk, from stumbling and falling to walking well. The early stages would probably be carried out in a virtual environment to prevent damage to the machine. Students of child development will see clear parallels here where caretakers create safe environments in which child can play and develop skills, and are gradually exposed to less protected worlds.

GOFAI stands for "Good Old Fashioned AI". It refers to when computers are pre-programmed with the aim of being able to cope with every future eventuality. What Tegmark describes is how a combination of deep learning and GOFAI gave rise to powerful and creative solutions. He saw this as similar to the combination of intuition and logic, a combination which will be familiar to many researchers in ethology – one tries to intuit an understanding of some phenomenon and then subject that to logical and empirical scrutiny. The parallels with child development are here too, children play and try out all sorts of behaviour, but also have built in (GOFAI) ways of processing input and controlling output. These interact with their specific exploration (finding out what something does) and with their diversive exploration (play - finding out what can be done with something) to develop new understandings and skills. (Berlyne,1960; Hutt, 1970; Plooij, 2003).

Tegmark then addresses four questions:
1. How can we make future AI systems more robust than today's, so that they do what we want without crashing, malfunctioning or getting hacked?
2. How can we update our legal systems to be more fair and efficient and to keep pace with the rapidly changing digital landscape?
3. How can we make weapons smarter and less prone to killing innocent civilians without triggering an out-of-control arms race in lethal autonomous weapons?
4. How can we grow our prosperity through automation without leaving people lacking income or purpose

1. Robustness. He offers a tour of AI's involvement in space exploration, finance, manufacturing, transportation, energy, healthcare and communication. Whilst lauding the net benefits of AI involvement, he discusses where things have gone wrong in these fields. The errors come under four headings: verification, validation, security and control which can be summarised as:

- verification – Did I build the system right?
- validation – Did I build the right system?
- security – Will it be threatened by hackers, malware, etc.?
- control – Is the interface with humans good enough?

2. Legal. Tegmark holds out the prospect of the administration of justice being faster and less biased if influenced or even largely controlled by AI, but that poses many threats not least that the "Robojudges" will be hacked by the accused having the resources to do so (but there is nothing new about human judges being corrupted in many parts of the world).

The content of the law would need to keep pace with AI advances. For instance, when a self-driving car crashes, who is responsible, its occupants, its owner, the manufacturer, or as one legal scholar has suggested, the car itself? In this last case the car would have to take out insurance and that would put pressure on manufacturers to produce ultra safe cars to avoid prohibitive premiums being required of the car's owner. Extending this, he asks, should computers have rights, say, to own property of even vote?

3. *Weapons.* Although AI controlled weapons offer some advantages over flesh and blood soldiery, Tegmark's position is that there should be international treaties to prohibit their development, like those treaties concerned with nuclear, chemical and biological weapons. A major problem is that if developed they are likely to be cheap and easily available and will thus fall into the hands of criminals, terrorists, despots, totalitarians and those bent on genocide. He points out there is not much difference between a drone that delivers an Amazon package and one that delivers a bomb.

4. *Occupations.* Just as machines have destroyed many manual jobs, and robots destroyed many skilled jobs, so now computer systems and AI, threaten many white collar jobs. They also risk increasing inequality and promote the interests of capital over labour (see also Christophers, 2020).

Given the takeover of jobs by computer systems, Tegmark gives three examples of questions about career areas he might advise young people to ask:

- Does it require interacting with people and using social intelligence?
- Does it involve creativity and coming up with clever solutions?
- Does it require working in an unpredictable environment?

If the answer is yes, computers are less likely to take over.

Will the AI revolution generate new jobs, as previous technological revolutions have done, confounding the Luddites. Tegmark is not so sanguine, and sees other solutions may be necessary . One solution is to have an universal basic income. Such schemes have been tried and found actually to increase economic activity as well as happiness (e.g., Bregman, 2014) confounding the sceptics. Universal basic income may go a long way to meeting financial needs, it does not address the issue of helping people find purpose, especially given that for many their job makes a significant contribution to their sense of purpose in life and their happiness, although for many others foregoing the meaningless drudgery of their jobs would be welcome. Tegmark uses "positive psychology" ideas to address this, noting that some of the factors which promote happiness and a sense of purpose are,

- a social network of friends and colleagues
- a healthy and virtuous lifestyle
- respect, self-esteem, self-efficacy and a pleasurable sense of "flow" stemming from doing something one is good at
- a sense of being needed and making a difference
- a sense of meaning from being part of and serving something larger than oneself

Maybe.

Although he doesn't mention it, a parallel to these ideas is in the Kontratiev cycles or waves, after the work of Nikolai Kontratiev (1925). Kontratiev argued that economic activity went in cycles triggered by new technologies. This was later brought up to date (in italics)[6] as follows, starting from:

- late 1700s: steam engine, cotton;
- early mid 1800s: railways, steel;
- late 1800s: electrical engineering, chemistry;
- early 1900s petrochemicals, cars;
- *late 1900s, information and communication technology;*
- *late 1900s: psychological health.*

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

The penultimate one clearly parallels Tegmark's topic, and the last one could be seen as partly driven by the existence of sufficient wealth to allow there to be a focus on psychological health, but also, more directly, by the consequences of AI taking over more human activities, and (perhaps) moving people further away from the lifestyle to which they are adapted, with the negative consequences on mental health. Tinbergen (1972) foresaw this issue in his 1972 Croonian lecture when he pointed out how the faster and accelerating pace of cultural evolution, compared to genetic evolution, takes us further away from the lifestyle to which we are genetically adapted. Human ethologists are in a good position to contribute to ways of addressing this issue in the Psychological Health wave.

Continuing this theme, Tegmark asks if we shall ever achieve Human Level AGI. He thinks it is possible and that this Artificial General Intelligence may surpass the intelligence of humans, and this would give them a selective advantage in the same way as he sees humans as having a selective advantage over others because they are "smarter". Biologists might want to question whether being "smarter" is the only way to get selective advantage.

Tegmark considers various scenarios where hyper intelligent machines could break out of human control and essentially take over. The breakout scenarios are considered in detail, but the details of what such a post takeover world would look like are more scarce, although miniature killer robots (slaughterbots) and hyper surveillance systems are mentioned. Tegmark's Future of Life Institute made a cautionary and disturbing video about slaughterbots[7].

Tegmark's next chapter is titled, "Aftermath: The Next 10,000 Years". In it he describes, and then critiques, possible future scenarios involving AI. As in other places in the book Tegmark necessarily talks in hypotheticals and generalities because he is imagining future scenarios. In many cases, logic and reasonable assumptions hold sway, at others one is left with a sense that jumps have been made in the narrative that are pretty arbitrary. This not a criticism of the book but simply pointing out that imagining future scenarios will inevitably take on the flavour of story-telling fiction. But this is consistent with his repeated and correct theme that such thinking is necessary to anticipate the possible implications of AI. As with all such creative imagining, much will turn out to be nonsense, but that does not invalidate the effort. This make the book both interesting but also hard going in places as one thinks, "OK but is this worth thinking about now, it is so fanciful?"

As a human ethologist I would have liked him to relate his various scenarios to social organisations that have existed already, and most which he describes have, and compare these to his scenarios. But Tegmark is a physicist by training, and although now an accomplished, humane and creative polymath, taking Magpie like from the biological and social sciences, as well a literature, movies and everyday culture, his hard physical science way of thinking shows through.

An example of this is his easy use of the concept of ethical, as if what was ethical was self evident. Any moral judgement can be broken down into two questions, "*What* is it good for?" working back to more and more fundamental human needs. When that regression is exhausted, one asks, "*Who* is it good for?". Here we touch on the potential conflict of interests between AI machines/ systems and humans, already mentioned. Tegmark's narrative, emphasising intelligence, is often about the "What?" question, he notes the potential super intelligence of AI, and by implication its value as a tool for humans. But then the problems which he rightly urges us to address, also embrace the "Who?" question. Crucially, to be a life form (Life 3.0) it will need to be motivated to maintain its survival and reproduction and this may well bring it into conflict with humans. One solution to this, already mentioned, is Tegmark's alignment of goals, as in cooperative human

[7] https://www.youtube.com/watch?v=9CO6M2HsoIA

relationships. But one need look no further than the old adage, “My enemy's enemy is my friend” (e.g. USSR and the Western Allies vs Nazi Germany, later changed to USSR vs the West), to see how fragile that solution is.

One of the scenarios is that the super intelligent AI acts as a “Protector God”, omniscient and omnipresent. He writes that many people might like this scenario because it is similar to current monotheistic religions, and if “someone asks the super intelligent AI “Does God exist?” after it's switched on, it could repeat a joke by Stephen Hawking and quip “It does now!” (page 178). I believe it was in the Brecht play on Galileo, that there was the line, uttered soto voce, something like, “if God did not exist, men would invent one”. In this scenario, men really have created God. One is reminded of the religious cults of creationism and intelligent design, but in the Protector God scenario there really is intelligent design - by super intelligent AI. Another scenario is “Enslaved God” in which AI has not “broken out” and remains a tool which humans use. But as he says, “The greater our breakout paranoia, the less AI-invented technology we can use.” (page 180). This is the familiar idea often applied to totalitarian societies, who stifle freedom of communication of ideas, and therefore, it is argued, creativity and adaptability to change, lest it leads to their loss of control.

After usefully discussing these scenarios, Tegmark moves into more rarefied speculation about the “next billion years and beyond”, thence he moves to “Goals “and then to “Consciousness” and the quality slips, as I discussed earlier.

One omission from the book is sustainability. Given the huge amounts of energy computer farms consume, will a human level intelligence AI consume so much energy as to be unsustainable? The energy consumption of biological brains may turn out to be more sustainable.

LAST WORD

Max Tegmark's book is particularly useful when it discusses AI and the possible benefits and threats which its development implies. It becomes much less coherent when moving into areas more familiar to human ethologists, psychologists, moral philosophers and logicians. This need not matter since the dangers and benefits of AI can be looked at without notions of Life 3.0, and consciousness. The positive aims of the book shine through especially in the epilogue which is a call the think about these issues and a create the future we decide we want. But the predictions and warnings would be greatly strengthened by the insights of biological areas like ethology.

REFERENCES

Berlyne, D.E. (1960). *McGraw-Hill series in psychology. Conflict, arousal, and curiosity.* McGraw-Hill Book Company. DOI

Bregman, R. (2014). *Utopia for Realists.* London: Bloomsbury

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies, 2*(3): 200–219.

Christophers, B. (2020). *Rentier Capitalism: Who Owns the Economy, and Who Pays for It?* London: Verso Books.

Dennett, D. C., (1987). Three Kinds of Intentional Psychology. In R. Boyd, P. Gasper, & J. D. Trout (Eds.), *The philosophy of science* (p. 631–649). The MIT Press. (Reprinted from R. Healey (Ed.), "Reduction, Time and Reality," New York: Cambridge University Press, 1981, 37-61.

Hutt, C. (1970). Specific and Diversive Exploration. In Reese, H.W. and Lipsett, L.P. (eds) *Advances in Child Development and Behavior, vol. 5*. New York: Academic Press. DOI

Kontratiev, N. (1925). *The Major Economic Cycles English version Nikolai Kondratieff (1984). Long Wave Cycle*. Guy Daniels. E P Dutton

Penrose, R. (1997). On understanding understanding. *International Studies in the Philosophy of Science 11*(1), 7-20. DOI

Plooij F. (2003). The Trilogy of Mind. In: Heimann, M. (ed). *Regression Periods in Human Infancy*. London. Lawrence Erlbaum Associates. DOI

Richer, J.M. (1986) *Consciousness is in the Mind of the Ascriber*. Paper presented at the 5th International Human Ethology Conference, Tutzing, West Germany.

Richer, J.M. (2016). Mentalistic and scientific stories about human behavior, biomimetic heuristics and psychology's confusions. *Human Ethology Bulletin, 31*(4), 15-33. DOI

Strawson, P.F. (1960). Freedom and Resentment. *Proceedings of the British Academy, 48*(1962).

Tinbergen N. (1972). The Croonian Lecture, 1972: Functional Ethology and the Human Sciences. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 182* (1069). DOI

Turing, AM (1950). Computing Machinery and Intelligence, *Mind, 59*(236). 433-460. DOI

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford, Blackwell