Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

28(1) – September 2018

The Value of Consciousness and Free Will in a Technological Dystopia

Allan McCay
University of Sydney Law School
University of Sydney Foundation Program
Centre for Agency, Values, and Ethics, Macquarie University

Amcc4688@uni.sydney.edu.au

Journal of Evolution and Technology - Vol. 28 Issue 1 – September 2018 - pgs 18–30

Abstract

Yuval Noah Harari warns of a very pessimistic future for our species: essentially, that it may be superseded by non-conscious Artificial Intelligence that can do anything we can and more. This assumes that we are physically instantiated algorithms that can be improved on in all respects. On such an assumption, our labor will become economically worthless once AI reaches a certain level. This picture, however, changes markedly if we accept the views of David Hodgson in respect of consciousness, free will, what he calls plausible reasoning, and the relationship among these. On Hodgson's account, there will always be valuable skills requiring a particular kind of judgment that are possessed by humans, but not by non-conscious algorithmic machines, however advanced.

Introduction

Are consciousness and free will worth anything? In this paper, I will focus on the value of consciousness and free will within economic systems of the future. I will do this by considering a recent book that envisages some dystopian future scenarios, in light of work at the intersection of philosophy of mind and the free will debate. The intention is to thereby consider philosophical debates about consciousness and free will, in a context that is some distance from traditional treatments of these topics – my context being one that is primarily concerned with the *economic* worth of human labor in a technological dystopia. In particular, I focus on the question of how humans will fare in economic competition with Artificial Intelligence (AI).

Homo Deus: A Brief History of Tomorrow (2016) is a popular book (written by Yuval Noah Harari, an academic historian[1] specializing in human history on a grand time scale) that is having a significant impact on popular debates about the future of technology and the future of humanity. According to Harari, we face a bleak future, and may become economically useless, as a result of being superseded by non-conscious AI and/or enhanced humans.

In developing his thesis, Harari attaches great significance to the concept of the algorithm. Throughout the book he works with a presupposition about the role of algorithms in human behavior that he claims is becoming a dogma[2] in scientific discussions and has also much wider impact on society. The presupposition is that humans are algorithmic choosers who lack free will. Whilst he seems concerned about this perspective on agency and regards it as important to expose it for critical evaluation (Harari 2016, 397), he also appears to find it plausible enough for it to be the view that he assumes in relation to his predictions about the economy. Harari's assumption about algorithmic human agency has great import for the way he envisages that we will engage in economic competition with AI and, in short, are likely to be outperformed. As a result of the book's focus on the relative merits of humans and AI in the workplace, Harari's argument provides a useful framework in which to consider an issue that is of great contemporary significance – that of the automation of work.

Whilst David Chalmers, for one, has considered the possibility of non-algorithmic AI (2015, 177), it is significant that *contemporary approaches to AI are algorithmic in nature.* Although recently there has been substantial progress in AI, for example in the branch of machine learning known as artificial neural networks (which is loosely inspired by the neural structure of the brain), it is important to note that, like earlier forms of AI, neural networks operate algorithmically. Although the algorithms that constitute a neural network can facilitate the machine learning required for impressive applications such as self-driving cars, the code that underpins such learning is still algorithmic.

For Harari, we have something in common with self-driving cars and other forms of AI insofar as the way we make decisions and navigate the world is also algorithmic. According to Harari, the risk to our economic value comes from a trajectory in the development of technology in which the algorithms underpinning AI outperform our more modest algorithmic agency in all ways that are economically useful.

However the philosophical work of David Hodgson challenges the analysis presented by Harari by suggesting a different picture of human agency. Hodgson also considers the role of algorithms, focusing on human decision-making, but in contrast to what is assumed in Harari's book Hodgson argues that human decisions are *not* merely algorithmic choices, but involve "plausible reasoning," which, according to him, gives a role for consciousness and a basis for free will.

As with Harari, David Hodgson (1939–2012) has an unusual background as a contributor to debates about free will and consciousness. After completing doctoral studies under the legal philosopher H.L.A. Hart, Hodgson went on to become a Judge of Appeal in the Supreme Court of New South Wales, Australia, and whilst on the bench he published philosophical work "at a rate that would be respectable for a full-time philosopher" (Levy 2013). In his later years, most of his work focused on free will and consciousness, culminating in the publication of his third and final monograph, *Rationality + Consciousness = Free Will* (2012) in which he argues for a libertarian form of free will that, he claims, could make people morally responsible and deserving of punishment for what they do. It is the view of free will and consciousness in this last book that I will pit against Harari's analysis, and I will consider the implications of accepting Hodgson's view for Harari's pessimistic predictions.

Why consider Hodgson's work in response to Harari? There are three reasons to focus on Hodgson. The first relates to Hodgson's discussion of algorithmic choice. Within the free will debate, Hodgson is perhaps the writer who engages in the most substantial discussion of algorithmic choice, and his comments on this issue raise questions about Harari's thesis. From time to time Hodgson compares[3] algorithmic and non-algorithmic agency, thereby providing useful insight for the analysis in this paper.

A second reason relates to Hodgson's focus on consciousness. One of the main questions in Harari's book, and this paper, relates to the value of consciousness in a technological dystopia. As noted by Levy (2013, 185), a significant feature of Hodgson's view is his emphasis on the role of consciousness in free will. Arguably, Hodgson's work has the most substantial discussion of

consciousness in contemporary debates about free will (aside from the empirical perspectives discussed by Levy (2017)).

A third reason relates to Hodgson's judicial background. Hodgson uses examples from his work as a judge to illustrate the kind of task for which consciousness and non-algorithmic reasoning are important. As Harari's thesis engages with the future of employment, it is useful to consider Hodgson's vocational (judicial) examples.

Although Nagel (2012) has described Hodgson's last book as a "persuasive account of the relationship between consciousness and free will," critics of the work have drawn attention to issues with various aspects of Hodgson's argument, ranging from those relating to his perspective on the laws of nature, to his analysis of evolution's role in the emergence of consciousness, and the ability of his account of free will to address concerns about control and luck, and to form a basis for retributive punishment (Carruthers 2012; Capes 2012; Earp 2013; Levy 2013; White 2013; Deery 2015; de Sio and Vincent 2015).[4] It is not possible to respond to these issues in a paper which addresses Harari's book, but it is worth noting that many of the criticisms do not relate to the *economic* argument presented here. For example, it is not necessary for me to assume that humans have a form of free will that might make them deserving of retributive punishment in order to show that their agency provides them with a capacity that is of value in a labor market.

My goal here is thus not to defend Hodgson's theory but to raise a serious possibility that what he calls plausible reasoning may give humans an economic advantage over non-conscious AI, and to suggest that this view should be borne in mind when considering the basis on which Harari's predictions are made, and the likelihood that they will come to fruition. I thus assume that Hodgson's views (or something like them) are defensible against pertinent criticisms, then *apply* them in the consideration of claims about future economic circumstances.

I begin by briefly saying something about the aims of Harari's book before setting out his claims about algorithms, consciousness, and free will. Later I outline various future scenarios that are developed in reliance on the claims. After that I use Hodgson's work to draw the claims into question, before revisiting the grim economic scenarios that Harari contemplates.

My Hodgson-inspired response to Harari suggests that, in the technological dystopias that Harari envisages, libertarian agents who can use a reasoning facility that derives from their consciousness might have a superior economic value to deterministic agents (at least of the algorithmic variety). Libertarian free will and consciousness may thus have a dollar value, and if we are indeed libertarian agents of the type described by Hodgson, our economic future might be less bleak than if we were deterministic agents, whose method of choosing is governed by rules.

Thus the following argument is methodologically unusual insofar as it challenges claims about the economy using a theory of free will and consciousness. This challenge to an economic forecast constitutes a novel application of such philosophical theorizing.

Harari on algorithms, consciousness, free will, and dystopia

In *Homo Deus: A Brief History of Tomorrow,* Harari sets out a number of possible future scenarios that might result from scientific and technological progress. He does not claim that any scenario is inevitable, and the book aims to identify some broad trends in the development of human societies in order to consider the merits of progressing one way or another. As a historian concerned with the macro trends in the trajectory of humans, Harari develops an analysis that is informed by a significant engagement with the past.

Harari's book could be said be to a contribution to the transhumanist tradition,[5] which considers ways in which the human condition might be transformed in the future. He claims that humans now have the capacity to control famine, plague, and war (even though this capacity is not always exercised due

to political failure). He anticipates that, having attained the ability (albeit not always the motivation) to control such phenomena, humans will turn their minds to the pursuit of immortality, happiness, and divinity (in the form of great powers) (Harari 2016, 21). According to Harari, the pursuit of these aims will be by way of biological engineering, cyborg engineering, and the creation of AI (2016, 43). He expects great technological advance in these areas.

This paper focuses primarily on the economic threat from AI, and in this connection it is worth noting that Harari envisages impressive forms of AI that lack consciousness.[6] He does not address the functionalist[7] account of consciousness, and his dismissal of conscious AI is not fully supported. However, he says, "[o]ver the past half-century there has been an immense advance in computer intelligence, but there has been exactly zero advance in computer consciousness. As far as we know, computers in 2016 are no more conscious than their prototypes in the 1950s" (Harari 2016, 311).

From this he appears to assume that "it doesn't seem that computers are *about to* gain consciousness" (Harari 2016, 311; emphasis added), and, although he does not specify a time frame for the emergence of the scenarios he warns of, his book is best understood as referring to a time in which AI is not conscious. He does not seem to suggest that AI will never become conscious, or that it will at some point, but he does not treat conscious AI as a serious possibility in respect of the economic scenarios he considers.

This is an important point for my purposes, as my focus is on the economic advantages of free will and consciousness over non-conscious AI and *my argument only relates to a time when AI is not conscious*. I am agnostic on the question of whether AI will at some point attain consciousness, and I will return to the possibility of conscious AI briefly toward the end of the paper.

Prior to considering the economic scenarios that Harari sketches, I will say something about the way that Harari conceives of human agency.

Algorithms and free will

As will be seen in the next section, an important premise in Harari's theorizing is that humans are algorithmic choosers who lack free will. This view of agency becomes the model of workforce agency that he assumes of humans when considering the relative values of humans and AI in the labor markets of the future. For Harari, algorithms are becoming "the single most important concept in our world" (Harari 2016, 83). Because of the importance of the concept for Harari (and for this paper), it is worth setting out how he defines it:

> An algorithm is a methodical set of steps that can be used to make calculations, resolve problems and reach decisions. An algorithm isn't a particular calculation, but the method followed when making the calculation. (Harari 2016, 83)

Why are algorithms so important? AI depends on algorithmic methods, by way of codified procedures, but another very important reason for the significance of algorithms is that, according to Harari, biologists have now concluded that *human beings are also algorithms* (Harari 2016, 85). He elaborates on this idea with a comparison between humans and coffee vending machines which, in algorithmic fashion, go though a series of steps to produce a cup of coffee. Harari sees humans as vastly more complex algorithmic agents: "The algorithms controlling vending machines work through mechanical gears and electric circuits. The algorithms controlling humans work through sensations, emotions and thoughts" (2016, 85).

In the analysis presented by Harari, "free will exists only in the imaginary stories we humans have invented" (Harari 2016, 283), and the algorithmic nature of decision-making has a role in our unfreedom. He says, therefore: "The algorithms constituting a human are not free. They are shaped by genes and environmental pressures, and take decisions either deterministically or randomly – but not freely" (Harari 2016, 328).

For the purposes of the discussion of Hodgson's work, it is useful to briefly contextualize Harari's comments on free will within the views prevalent in the scholarly debate. Much of contemporary philosophical debate about free will is focused on the compatibility of free will, moral responsibility, and/or blameworthiness with an indeterministic or deterministic picture of the world. Within that vast debate, some are compatibilists. These are philosophers who think that at least some of the following concepts are compatible with determinism: praise, blame, and deserved reward or punishment. Other philosophers think that concepts such as these presuppose libertarian free will. According to libertarians in the free will debate, praise, blame, and desert presuppose an *undetermined* form of choice, which is not random but controlled by the agent.

Whilst Harari is not primarily concerned with blame and punishment, what he refers to as free will is *libertarian free will.* Harari conceives of freely willed decisions as being "neither deterministic nor random" (Harari 2016, 281). For the purposes of his theorizing, Harari assumes that humans do in fact lack free will, and he suggests that scientific and technological advance means that people in general will lose their belief in free will:

> Doubting free will is not just a philosophical exercise. It has practical implications. If organisms indeed lack free will, it implies that we can manipulate and even control their desires using drugs, genetic engineering or direct brain stimulation. (Harari 2016, 286)

Thus, according to the view of agency that Harari works with, we are algorithmic choosers who lack free will, and who may be subject to various forms of manipulation based on new understandings of our unfree decision-making processes. But how does this view of agency figure in Harari's predictive economics?

Dystopias

Harari identifies three possible dystopian scenarios that might emerge from technological advance. One sees members of the species *Homo sapiens* completely losing their economic value, another involves them becoming managed by AI, and the third envisages the rise of technologically enhanced superhumans.

In describing the first scenario, Harari argues that the link between intelligence and consciousness is breaking (Harari 2016, 311). Humans have general intelligence and consciousness, but he envisages the rise of AI that is not conscious but may still outperform us. The significance of algorithms to Harari's thinking emerges in the following passage:

> The idea that humans will always have a unique ability beyond the reach of non-conscious algorithms is just wishful thinking. The current scientific answer to this pipe dream can be summarised in three simple principles:
>
> 1. Organisms are algorithms. Every animal – including *Homo sapiens* – is an assemblage of organic algorithms shaped by natural selection over millions of years of evolution.
>
> 2. Algorithmic calculations are not affected by the materials from which the calculator is built. Whether an abacus is made of wood, iron, or plastic, two plus two beads equals four beads.
>
> 3. Hence there is no reason to think that organic substrates can do things that non-organic algorithms will never be able to replicate or surpass. (Harari 2016, 319)

Thus, Harari assumes that we are algorithms, our organic substrate is not significant, and there is no reason to think that an algorithm lacking in consciousness, and operating on a non-organic substrate,

will not be able operate in the same way as us or better.[8] Thus non-conscious AI may one day equal or surpass us in intelligence. For Harari, the full import of this claim comes from a consideration of the effects on employment. Because AI will equal or surpass us, it will make us economically useless.

In relation to the second scenario, management of humans by algorithms, Harari claims that artificially intelligent algorithms will have the capacity to understand[9] us better than we understand ourselves. Non-organic algorithms, such as those provided by large technology companies, may come to monitor our bodies as we interact with the world, learning about what we like and dislike, and what stresses us or relaxes us. They may also learn our interests by monitoring what we read.

Accordingly many people will transfer authority to the external algorithms, and let them make decisions. People will no longer see themselves as being autonomous and in charge of their lives (Harari 2016, 329). Even decisions such as choice of partner will be made by artificially intelligent algorithms, since they will be regarded as more competent (Harari 2016, 337). The upshot of this transference of authority is that that humans might ultimately be managed by a large network of algorithms. Thus, even if humans retain some usefulness within the system, they will defer to these artificially intelligent algorithmic choosers.

The third scenario Harari envisages involves a split in humanity, in which some humans are upgraded through technological progress, and they make the significant decisions for society, whereas other humans are subordinate to these superhumans (and also to non-human algorithms) and are regarded as useless.

But are we algorithms?

Harari concludes by posing a series of questions, which he says he hopes will engage the reader long after finishing the book. One of the questions he asks is "are organisms really just algorithms?" (Harari 2016, 397). This important question, which will be addressed in the next part of this paper, is of great significance to his thesis. Indeed, earlier in the book he acknowledges that computers might have inherent limitations, compared to organisms, if the latter are not algorithmic: "If organisms function in an inherently different way to algorithms – then computers may work wonders in other fields, but they will not be able to understand us and direct our life, and they will certainly be incapable of merging with us" (Harari 2016, 345).

Before challenging Harari's predictive project, I will state the premises used in the development of his thesis that are to be questioned. These are as follows:

1. Human choice is algorithmic

2. Humans do not have libertarian free will

Challenging the algorithmic claim: Hodgson on consciousness and free will

If Hodgson is right then both of these premises are false. *Human choice is not algorithmic and we have libertarian free will*. According to him, it is our consciousness, together with a form of rationality that is unavailable to non-conscious AI, that gives us libertarian free will. Acceptance of Hodgson's view would lead to a rejection of some of Harari's assumptions about the agency of workers, and it would show a need to reconsider the scenarios envisaged.

Before considering its implications for Harari's predictions, I will summarize Hodgson's account of free will. There are a number of steps in the development of the theory and it is not easy to condense it into short form, but I will now give an outline prior to applying it in the next section.

Hodgson claims there are two types of reasoning that humans can perform: algorithmic reasoning and plausible reasoning. According to Hodgson, humans are able to engage in formal reasoning (such

reasoning takes place in logic and mathematics). In these domains, the conclusions are entailed by the premises and rules. Hodgson also calls this algorithmic reasoning, and there is a mechanical sense to such reasoning which allows computers to operate this way. Hodgson and Harari could agree that computers work algorithmically, but importantly, whilst Hodgson would say that humans can operate algorithmically, contrary to the view of human agency assumed by Harari, Hodgson does not suggest that this is the *only* way that humans operate.

According to Hodgson, humans also engage in reasoning in which the conclusions are not entailed by the reasoning process, because the reasons are inconclusive. This form of reasoning, he calls plausible reasoning (Hodgson 2012, 37–53).

A form of plausible reasoning takes place when humans are weighing incommensurables. Thus if one is deciding whether to honor a promise to help a friend, or to go for dinner with a person one finds attractive, the competing reasons are of a different kind (duty and desire) and thus are incommensurable. The reasons do not lead to conclusions through entailment, and judgment is required. This involves reasoning about what to *do*, but Hodgson argues that reasoning about what to *believe* is also a form of plausible reasoning. An example of this comes from Hodgson's judicial experience in which he describes a judge trying to decide what to believe after hearing inconclusive evidence:

> Very often, there are contemporary documents that to some extent record what happened, and there is sworn oral evidence given by witnesses in court. There may be conflicts between oral evidence given by different witnesses, and some or all of this evidence may conflict to a greater or lesser extent with contemporary records. Opposing accounts of what people did may appear to conform to how one would expect people to behave in the circumstances in question, or to deviate to a greater or lesser extent from such expectations. The behavior of the witnesses in giving their evidence may give an impression of honesty or dishonesty, or of good or poor memory. Evidentiary conflicts of these kinds are not, and I say cannot be, resolved by explicit formal reasoning. The conflicting considerations are inconclusive and of different kinds, and there is no way in which a judge can overtly apply precise rules of formal reasoning to them in such a way as to determine a conclusion. And yet, I contend, reasonable albeit fallible decisions are made. (Hodgson 2012, 38)

Hodgson maintains that plausible reasoning is an important aspect of human rationality. For example, he also argues that scientists rely on plausible reasoning when they generate hypotheses, consider suitable experiment methodology, and provisionally select unrefuted hypotheses (2012, 39). If Hodgson is right, then it would seem that plausible reasoning is something that would be problematic for algorithmic agents. According to him, humans are able to make decisions that involve the weighing of incommensurables, and this is taken as evidence in support of his thesis that the decisions of humans do not always proceed algorithmically. He claims that this leaves space for an undetermined form of agency, where the reasons do not determine the outcome of a decision.

Hodgson accepts the possibility that, underlying the level of conscious reasoning, there may be an unconscious algorithmic process which accounts for decisions that, at the conscious level, involve plausible reasoning. However, he does not take this to be likely, as, for evolution to have led to consciousness, Hodgson claims, consciousness must have had a role. Evolution does appear to have favored consciousness and so we can assume it does have a role. But what could that role be?

According to Hodgson, humans can engage in plausible reasoning and make reasonable decisions. In such cases, they do this by consciously responding to gestalts (2012, 79–113). He argues that agents experience sets of circumstances as gestalts, in which the whole experience is greater than the sum of the parts. This experience of the circumstances as a whole, rather than merely a series of component parts, influences agents' decisions. Hodgson claims such appraisal could provide a role for consciousness. Its role would be to appraise a gestalt in a way that is a positive contribution to decision making.

So on this view, the appraisal of the gestalt contributes to the decision-making process. As the particular features of any whole set of circumstances that require a decision are unique, Hodgson argues that no law could dictate a response to such a gestalt, as laws apply to classes of things. Although one could perhaps respond algorithmically to component aspects of a set of circumstances that are factors in a decision, one could not respond algorithmically to the gestalt, to the experience of the *whole* situation (such as that of judge described earlier), due to its uniqueness.

So according to Hodgson, laws need to apply to classes of things, and so any response to the uniqueness of the gestalt could not be governed by a law. As humans can indeed respond to gestalts, Hodgson maintains, it seems unlikely that they are governed by laws that conclusively dictate action and this makes a deterministic view of agency less plausible, and makes it more plausible that we are libertarian agents. On his view, we have the capacity to make decisions that are not wholly determined by the way we are and the laws of nature.

Hodgson admits he is making a claim that might be falsified by science, but he argues that science as it currently stands does not suggest this account is wrong. In doing this, he points to a scientific worldview – that of quantum physics. This view suggests the possibility of indeterministic causation in which causes do not determine their effects but merely raise the probability of a certain effect. If this interpretation of quantum physics is correct, then determinism is false and the world is indeterministic. Thus, like Harari, Hodgson accepts that scientific understanding is of significance for the free will debate, but he claims his own libertarian view is not refuted by science.

The following seem to be implications of Hodgson's work that are particularly salient for a response to Harari:

1. We have a capacity for plausible reasoning

2. This capacity derives from our consciousness

3. We are indeterministic choosers as we have libertarian free will

The value of consciousness and free will in a technological dystopia

Will we be useless in the coming technological dystopia? If we retain some limited usefulness, will artificially intelligent algorithms find us easy to manage? If Hodgson is right, then perhaps the answer is "No" to both questions or at least to the first.

On Hodgson's view, our consciousness gives us a capacity that non-conscious AI cannot attain – we are plausible reasoners. This enables us to perform certain sorts of jobs that these algorithms cannot do, or at least cannot do as well (such as the judging and the scientific tasks described earlier). However, the need for plausible reasoning seems to be widespread in the workforce. Consider the work of an architect, who must balance aesthetic considerations against considerations pertaining to the bearing of load in a structure. Or the many occupations that require a weighing of ethical considerations against commercial factors.[10] If one were to accept Hodgson's view, perhaps it is unsurprising that the job of social worker, with its many competing considerations, has been regarded as one that is at very low risk of automation (Frey and Osborne 2017). As plausible reasoners, we may be significantly more useful in the labor markets of the future than Harari envisages, as the need for plausible reasoning seems to be fairly pervasive in many occupations.

Even if artificial algorithmic agents surpass us in many intelligence domains, and ultimately dominate us, so long as they lack consciousness they might find us useful for tasks involving plausible reasoning (if such tasks enable them to pursue their aims).[11] Assuming the inability to engage in plausible reasoning does not prevent the rise of the algorithms, but they might need us for the many kinds of work that require the weighing of incommensurables. To the extent that plausible reasoning

is useful in the formation of beliefs, we might be used as epistemic workers – helping the artificially intelligent algorithms to know about the world in ways that they cannot otherwise. To the extent that these algorithms aim to act so as to accomplish tasks that involve the weighing of incommensurables, they might need to rely on our advice as to how they should act, or perhaps delegate that work to us. Rather than our becoming useless, it seems that the algorithms may need us in a novel division of labor.

Will the algorithms successfully manage us? Perhaps the economic value of conscious plausible reasoning goes some way in explaining why the algorithms might be motivated to engage in this kind of management – they might need us for certain kinds of work that are instrumental to their aims, and need to manage us to ensure that we do it.

In relation to the management scenario, Harari envisages algorithmic forms of AI that have knowledge about us that exceeds our self-knowledge.[12] According to Harari, because the algorithms know us so well, we will see them as better decision-makers and will entrust our decision-making to them. Even if we thought it was prudent[13] to trust the algorithms, how far would we go in entrusting our decisions to them? That is an interesting empirical question and perhaps we should not assume that people will act in a way that they deem prudent, and thus defer to algorithms. But from now on, I will assume this trust.

Harari claims that the algorithms may use their knowledge of us in order to manipulate us, but he does not clearly specify the means of manipulation. He appears to envisage that they will use their knowledge to give us advice that, based on their intimate knowledge of us, we are likely to follow (rather than use a manipulative method such as direct brain stimulation) (Harari 2016, 334–46). This seems to involve the algorithms using their knowledge to predict how we will react to their advice.

If Hodgson is right about libertarian free will, there will be a limit to the extent to which our choices can be predicted, no matter how much is known about us as agents. Libertarian choices are undetermined, *even by the agent themselves,* and to know everything about how an agent is prior to a choice is not to know how they will choose.

Similarly to another libertarian, Kane (1998, 74–77), Hodgson envisages the possibility of undetermined self-forming actions, in which agents make choices that have the effect of changing their characters (2012, 175). Thus, after a particular self-forming action, some of what the artificially intelligent manager knew about the agent prior to the choice might no longer be true of the post-choice agent. So, perhaps the managing algorithm would now need to wait and get more information about the post-choice agent before it could be said that they know that transformed agent. A similar problem would recur after the next self-forming action. Depending on how frequently self-forming actions take place, this might be problematic for algorithmic managers, and of course this issue would only compound any difficulties they had in reassessing the human agent after unforeseen changes brought about by the agent's environment.

Perhaps the algorithms would not need to know *with certainty* how humans will act in order to manage them effectively, and maybe fairly good predictability would suffice for their purposes. Indeed, today some members of the species *Homo sapiens* are able to manage others, despite the issues outlined in this section. Furthermore, there have been concerns expressed about manipulation by AI that already occurs (Yearsley 2017). That said, if Hodgson is right, our libertarian free will would appear to put some limitations on the algorithms' management capabilities. It is unclear, however, how much of an impediment this would be.

It is worth noting that this impediment might not be altogether a good thing for *Homo sapiens*, as the management issue might detract from our economic value in the employment market. However, so long as members of our species are being managed, it seems we are being regarded as in some way useful.

If one were to accept Hodgson's reasoning, and to conclude that there is something useful about plausible reasoning, this might be a reason to regard Harari's third scenario as a little more plausible. If *Homo deus* is an enhanced version of *Homo sapiens*, then it will presumably retain the capacity for plausible reasoning, and possess that capacity together with other enhanced capacities. This might confer an advantage over non-conscious AI, at least in some situations. *Homo deus* might still value the plausible reasoning of *Homo sapiens*, and use members of that subordinate species for work that requires this capacity.

Hodgson's work seems to suggest that there may be something that conscious agents can do that non-conscious algorithmic AI, *in principle*, cannot do. This, of course, is a strong claim. However, even if one were not convinced that it is, in principle, impossible for a non-conscious algorithmic AI to achieve something as economically useful as the capacity for plausible reasoning, one might still see merit in Hodgson's arguments. One might take the view that Hodgson's approach at least identifies issues, such as the rational resolution of incommensurables, that will be very difficult hurdles for those engaged in the production of AI. This then has consequences for Harari's three scenarios. It might mean that the worst of the economic effects on *Homo sapiens* are likely to be delayed until the hurdles have been cleared.

It must be acknowledged that the arguments in this paper are predicated on the assumption that Harari is right in assuming that artificial agents will operate algorithmically and are not likely to become conscious in a time frame that is worth considering. So if conscious AI that could engage in plausible reasoning were to be developed, humans might then become "useless" in the manner envisaged by Harari. Any reassurance provided from this paper relates only to a period in which AI is not conscious, and so the argument will be of greater comfort to a person who thinks that non-conscious AI is impossible, or very difficult to achieve, than to one who thinks its development is imminent.

A final point is worth noting prior to concluding. As has been argued, in each of the scenarios Hodgson's form of libertarian agency appears to confer an economic advantage, at least over determined agents of the algorithmic variety. Whilst economists have turned their mind to the issues of free will (Altman 2006), as far as I am aware, this is the first paper to argue for the economic benefit to the agent of possessing a form of libertarian free will. It consequently represents a novel (and worldly) step in debates about free will.

Conclusion

Through grim speculation on the way that the algorithmic nature of our agency may negatively impact on our ability to compete with AI, Harari raises the question of whether consciousness and free will are worth anything. On Harari's analysis free will is an illusion, consciousness has little economic value, and our future looks bleak.

However, if Hodgson's arguments about consciousness, plausible reasoning and free will are sound, then we are not algorithmic choosers. This might not avert some of the dystopian future scenarios that Harari considers, but the conscious exercise of libertarian free will, by way of plausible reasoning, might mitigate some of the negative consequences. In particular, consciousness might give us a valuable capacity that helps us decide what to believe, and what to do, when faced with incommensurables. Crucially, this would be a capacity that non-conscious AI lacks.

Homo sapiens might therefore retain some value in a technological dystopia because of the continuing value of consciousness and free will. Hitherto, those engaged in the free will debate have generally addressed other, perhaps loftier, questions about freedom, but if one now asks what turns on the issue of whether we have free will, on the view presented here the answer might be: our livelihoods.

Acknowledgments

I am grateful to Paul Dower, Neil Levy, David Baker, Jesse Cunningham, Oisin Deery, Glenn Carruthers, Paul Ham, Mike Bain, Gary Edmond and Meredith Burgmann, Andrew Haywards, Sam Wyper, Kevin Walton, and Michele Loi for their useful comments about all or parts of drafts of the paper. I am also grateful to David Christian and Elle Hrobat from the Big History Institute, at Macquarie University, for their help and to Russell Blackford and the anonymous reviewers for their constructive suggestions.

Notes

1. Harari also publishes on military history.

2. The word "dogma" has a religious association and this does not seem to be unintentional. Harari sees the algorithmic view of human agency as part of an emerging religious worldview which he calls "Dataism" in which the universe is thought to consist of "data flows, and the value of any phenomenon or entity is determined by its contribution to data processing" ((Harari 2016, 367). Harari calls for a questioning of this worldview (Harari 2016, 394).

3. See Hodgson (2012, 82–94) for an example.

4. For an edited collection with a further set of responses to Hodgson's last book see McCay and Sevel (forthcoming).

5. See Bostrom (2005) for an overview of this tradition.

6. His assumption has been challenged by Gray (2016).

7. As noted by Chalmers, functionalist theories of consciousness maintain that "what matters to consciousness is not biological makeup but causal structure and causal role, so that a nonbiological system can be conscious as long as it is organized correctly" (2015, 202). For Harari, we are both algorithms and conscious. Having accepted that biological algorithmic systems such as humans can be conscious, perhaps Harari should take more seriously the possibility that highly sophisticated non-biological algorithmic systems could be organized in such a way as to become conscious if he does not have a reason to reject the functionalist view of consciousness.

8. Damasio (2016) has objected to Harari's assumption about the insignificance of the substrate.

9. As Harari does not envisage conscious AI, presumably he does not envisage conscious understanding, knowing, or learning by AI devices.

10. I am grateful to one of the anonymous reviewers for suggesting these examples.

11. As Bostrom has noted, there are very significant difficulties in predicting the aims of any future artificial intelligence (2014, 127–39).

12. For an account of some practical computational problems related to such technology see First (2017).

13. First (2017) has persuasively argued that it may not be prudent to put too much trust in algorithms, bearing in mind that the corporate creators of the algorithms may well create them with their own interests in mind, rather than the interests of the agent for whom they "facilitate" choice.

References

Altman, M. 2006. Human agency and free will: Choice and determinism in economics. *International Journal of Social Economics* 33(10): 677–97.

Bostrom, N. 2005. A history of transhumanist thought. *Journal of Evolution and Technology* 14(1): 1–25.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Capes, J.A. 2012. Review of *Rationality + consciousness = free will. Notre Dame Philosophical Reviews*, July 6.
https://ndpr.nd.edu/news/rationality-consciousness-free-will/ (accessed September 5, 2018).

Carruthers. G. 2012. Review: *Rationality + consciousness = free will* by David Hodgson. *Metapsychology* 16(43), October 23, 2012.
http://metapsychology.mentalhelp.net/poc/view_doc.php?type=book&id=6670&cn=394
(accessed September 5, 2018.

Chalmers, D.J. 2015. The singularity: A philosophical analysis. In *Science fiction and philosophy: From time travel to superintelligence*, ed. S. Schneider, 171–224. Oxford: Blackwell.

Damasio, A. 2016. We must not accept an algorithmic account of humanity. *Huffington Post*, June 28 (updated December 6, 2017).
http://www.huffingtonpost.com/antonio-damasio/algorithmic-human-life_b_10699712.html (accessed September 5, 2018).

Deery, O. 2015. *Rationality + consciousness = free will*, by David Hodgson. *Mind* 124(493): 347–51.

de Sio, F.S., and Vincent, N.A. 2015. *Rationality + consciousness = free will* by David Hodgson. *Criminal Law and Philosophy* 9(4): 633–44 .

Earp, B.D. 2013. Does rationality + consciousness = free will? *Journal of Consciousness Studies* 20: 248–53.

First, D. 2017. Will big data algorithms dismantle the foundations of liberalism? *AI & Society*, June 6. DOI https://doi.org/10.1007/s00146-017-0733-4.

Frey, C.B., and Osborne, M.A. 2017. The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change* 114: 254–80.

Gray, J. 2016. Humanity Mk 2: Why the future of humanity will be just as purposeless as the past. *New Statesman*, October 13.
https://www.newstatesman.com/culture/books/2016/10/humanity-mk-ii-why-future-humanity-will-be-just-purposeless-past (accessed September 5, 2018).

Harari, Y.N. 2016. *Homo deus: A brief history of tomorrow*. London: Harvill Secker.

Hodgson, D. 2012. *Rationality + consciousness = free will*. New York: Oxford University Press.

Kane, R. 1998. *The significance of free will.* Oxford: Oxford University Press.

Levy, N. 2013. Hodgson, David, *Rationality + consciousness = free will*. *Australasian Journal of Philosophy* 91(1): 183–92.

Levy, N. 2017. Empirical perspectives on consciousness and its relationship to free will and moral responsibility. In *The Routledge companion to free will*, eds. K. Timpe, M. Griffith, and N. Levy, 434–44. New York and Oxford: Routledge.

McCay, A., and Sevel, M., ed. (forthcoming). *Free will and the law: New perspectives*. Oxford: Routledge.

Nagel, T. 2012. *Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false.* New York: Oxford University Press.

White, V.A., 2013. David Hodgson, *Rationality + consciousness = free will*. *Philosophy in Review* 33(2): 126–28.

Yearsley, L. 2017. We need to talk about the power of AI to manipulate humans. *MIT Technology Review*, June 5. https://www.technologyreview.com/s/608036/we-need-to-talk-about-the-power-of-ai-to-manipulate-humans/ (accessed September 5, 2018).