Decision theory presupposes free will

Christian List*

July/August 2024

Abstract: This paper argues that decision theory presupposes free will. Although decision theorists seldom acknowledge this, the way decision theory represents, explains, or rationalizes choice behaviour acquires its intended interpretation only under the assumption that decision-makers are agents capable of making free choices between alternative possibilities. Without that assumption, both normative and descriptive decision theory, including the revealed-preference paradigm, would have to be reinterpreted in implausible ways. The hypothesis that decision-makers have free will is therefore explanatorily indispensable for decision theory. If we regard explanatory indispensability as an indicator of reality in science, decision theorists should embrace the idea of free will.

1. Introduction

The aim of this paper is to argue that decision theory, of the kind studied in economics and neighbouring fields, presupposes that decision-makers have free will. Free will, roughly speaking, is the ability to choose and control one's own actions. The idea of free will is familiar from commonsense reasoning about human agency, and it is at the core of society's practices of holding people responsible for their actions, both in morality and in the law (for overviews, see Kane 2011 and Fischer, Kane, Pereboom, and Vargas 2007). When I chose tea over coffee this morning, for example, it seemed that my choice was free, and that I could have chosen otherwise. People are thought to be responsible only for actions they did out of their own free will. However, since at least the 1970s many scientists, across several fields, have argued that free will is an illusion (e.g., Pereboom 2001, Gazzaniga 2011, Harris 2012, Eagleman 2015, Sapolsky 2023). Free will, they say, has no place in a scientific worldview. The biologist Jerry Coyne (2014), for instance, writes: "Our thoughts and actions are the outputs of a computer made of meat – our brain". Since this computer "must obey the laws of physics", Coyne continues, "[o]ur choices … must also obey those laws". This, he suggests, undermines the idea of free will: "that our lives comprise a series of decisions in which *we could have chosen otherwise*." The economist and decision theorist Itzhak Gilboa (2007, p. 1) similarly suggests that "free will is a highly problematic concept, even if no form of determinism is assumed", and concludes that "[f]ree will is an illusion, requiring that one would suspend knowledge about oneself". However, he thinks that "[t]his illusion is … essential to rational decision making".

In this paper, I will argue that, from a decision-theoretic perspective, we should not consider free will an illusion. Rather, decision theory relies on the hypothesis that decision-makers have free will, in a way that decision theorists seldom acknowledge. Furthermore, I will suggest, there are good scientific reasons for tentatively accepting this hypothesis. My argument for realism about free will in decision theory builds on a recently proposed response to the scientific challenge for free will (List 2014, 2019, 2023). This response asserts that, far from being out of place in science, the idea of free will is well-supported by the sciences of human behaviour, insofar as many explanations in those sciences presuppose that humans have free will. I will show that an analysis of decision theory strongly supports this point.

---

* Munich Center for Mathematical Philosophy, LMU Munich, 80539 München, Germany. This paper was first presented as an invited talk at the 2024 D-TEA Workshop in Paris, June 2024. I am grateful to the participants and especially Itzhak Gilboa and Larry Samuelson for helpful comments and discussion.

2. Free will

What is free will?[1] I will assume that free will requires three things: intentional agency, alternative possibilities among which one can choose, and causal control over the resulting actions. In more detail:

> **Intentional agency:** To be a bearer of free will, one must be an intentional agent, i.e., an entity that interacts with its environment in an intelligible, goal-directed manner.
>
> **Alternative possibilities:** To be a bearer of free will, one must have alternative courses of action to choose from.
>
> **Causal control:** To be a bearer of free will, one must have relevant control over one's actions, in the sense that those actions must be appropriately caused by one's intentional mental states ("mental causation"), not just by some sub-intentional, physical processes.

Different philosophical theories of free will differ in how important they take each requirement to be and how they explicate their details (again cf. Kane 2011). Some, in particular, suggest that not all three requirements are needed for free will. So-called compatibilist theories, for instance, redefine and weaken the requirements such that even someone who never faces any real choices between alternative possibilities could still count as having free will. For instance, someone might count as having free will if his or her behaviour is appropriately supported by his or her intentions, independently of whether this person could ever have acted differently. For so-called incompatibilist theories, by contrast, this would not be enough. For them, free will requires real alternative possibilities to choose from: "forks in the road" at each point of choice. (Technically, *incompatibilism* is the view that free will is incompatible with pre-determination of one's actions.) For my argument, I do not want to rely on any watering down of the idea of free will, and so I will assume that the three requirements – intentional agency, alternative possibilities, and causal control – are jointly necessary and sufficient for free will.[2]

According to free-will sceptics, it is hard to justify the view that human beings meet these requirements. Some sceptics think that our law-governed physical universe leaves no room for real intentional agency, and that intentional agency is an outdated idea from folk psychology (see, e.g., P. M. Churchland 1981 and P.S. Churchland 1986). Others think that the physical universe does not leave any room for real choices between alternative possibilities, because of deterministic laws of nature. Our actions, they say, are always pre-determined by prior physical conditions (see, e.g., van Inwagen 1975 and Sapolsky 2023). Still others think that our conscious

---

[1] In developing my argument, I build on the analysis of free will and the strategy of responding to scientific challenges for free will in List (2014, 2019, 2023). The framing and general strategy outlined in Sections 2 and 3 particularly draw on that earlier work. I also build on List and Rabinowicz (2014).

[2] Some people might use the term "freedom of action" for what I have called "free will", and reserve the term "free will" for something even more demanding. Free will, on such a more demanding understanding, might require control not only over one's choices or actions but also over all of their relevant preconditions, including one's "will-formation" – in the limit, the entire history of one's preference formation. We might call this "ultimate freedom". But this is clearly unrealistically demanding, and it's not the notion of free will I want to consider here.

intentions are mere "epiphenomena"; our actions are caused by sub-conscious brain activity; our intentions are just subjectively experienced byproducts (Libet et al. 1983, Kim 1998).[3]

Note, however, that this kind of free-will scepticism tends to view human beings primarily as biophysical systems: conglomerates of interacting cells or networks of neurons exchanging electrical signals. The sceptics are right that the notions of intentional agency, choice between alternative possibilities, and mental causation do not fit easily into this way of representing human beings. Notions such as agency, choice, and control are largely absent from the physical sciences and even from much of neuroscience. However, from the fact that these notions cannot be found in those sciences, we cannot infer that the relevant phenomena are unreal. Although the human organism can be viewed as a biophysical system *at some level*, this is arguably the wrong level of analysis for explaining more complex human behaviours.

Here is an analogy: the physical sciences also do not speak about organisms, universities, or economies, and yet we would not conclude that those phenomena are unreal. They are higher-level phenomena, which cannot be described using the concepts and categories of the physical sciences alone, but they are nonetheless real. There are many phenomena that we consider perfectly real, even though they are nowhere to be seen at the level of physics or of any other low-level science. Intentional agency itself is arguably one such phenomenon.

These considerations suggest that the physical sciences are not the right place to settle the question of whether human beings have free will. To corroborate or reject the hypothesis that humans have free will, we must look at the sciences of human behaviour and especially the sciences of human decision-making.

3. Free will in the human and social sciences

To introduce my argument that decision theory presupposes free will, I will first outline a general argument for the claim that many explanations in the human and social sciences presuppose free will. I will then turn to decision theory in particular. My general claim will be this:

> **The explanatory indispensability of the free-will postulate:** Many explanations of human behaviour depend on the postulate that humans are intentional agents, with alternative possibilities to choose from, and relevant control over their actions.[4]

Specifically, I will argue that postulating intentional agency and alternative possibilities is explanatorily indispensable in many of the human and social sciences. Given space constraints, I will not discuss causal control here, which raises more technical issues that are beyond the scope of this paper; I have discussed it elsewhere (e.g., List and Menzies 2009, List 2019).

---

[3] Kim's (1998) argument targets any form of mental causation that takes mental causes to be distinct from merely physical causes at the level of the brain.

[4] In List (2014, p. 169), this claim was made as follows: "free will, in the sense of being able to choose from more than one option, is explanatorily indispensable in our best scientific theories of agency" (p. 169).

Let me first consider intentional agency. Why is it explanatorily indispensable to postulate that people are intentional agents? The key point to note is that although the various human and social sciences, which range from anthropology and psychology to political science, sociology, and economics, differ significantly in their methods and approaches, they have one important feature in common: they all offer *intentional explanations* of human behaviour. That is, they understand humans as intentional agents: beings with goals and beliefs and a capacity for perception and thought, who act in at least approximately intelligible ways.

Consider the kinds of questions that social scientists, including economists, might seek to answer:

- Why, and under what conditions, do people engage in religious practices?
- Why do consumers make the choices they do?
- Why, and under what conditions, do people cooperate with others even when they would benefit from defecting?
- Why do some people vote for populist parties while others do not?

We would have no idea where to begin if we tried to answer these questions without viewing people as intentional agents. Contrast intentional explanations with dynamical explanations, of the sort given in the physical sciences. In a *dynamical explanation*, as in a Newtonian model of the solar system, we describe deterministic or stochastic state-change rules by which a system changes its state over time. Future states of the solar system, for instance, can be described as a function of its present state; future states of a weather system can be described using random variables that depend on the present state. Dynamical explanations are purely mechanistic: they do not appeal to anything like choice, intention, preference, belief, rationality etc. Their logic is fundamentally different from that of an intentional explanation, in which a decision-maker is described as making intelligible choices between different possible courses of action: the decision-maker's "options". The moon, Mars, or the weather make no choices. They have no "options". Their state evolves deterministically or stochastically over time.

Non-intentional explanations, such as physical or neuroscientific ones, aren't remotely capable of accounting for human behaviour in its full breadth and flexibility. They may *complement* intentional explanations, by shedding light on some of the biological and neural implementation mechanisms of intentional agency, telling us, for instance, how the brain realizes certain cognitive functions. But they do not *replace* intentional explanations. Postulating intentional agency remains explanatorily indispensable. The philosopher Daniel Dennett (1987) made a similar point when he noted that, to understand human behaviour, we must often take an "intentional stance" towards people, i.e., view them as intentional agents. Taking a "physical stance" and looking for dynamical explanations may work in the case of purely physical systems such as the solar system but will be inadequate if we wish to answer the kinds of questions about human behaviour that social scientists usually ask. It seems absurd to try to explain religious practices, consumer choices, human cooperation, or voting behaviour without acknowledging that people have intentional agency.

Now, once we explain people's behaviour by viewing them as intentional agents, the postulate that they have alternative possibilities to choose from turns out to be indispensable too. Suppose we wish to explain why a particular agent acts in such-and-such a way, or why and when that agent will act in one way rather than another. When we answer such questions, we typically hypothesize the following, and we then test the resulting three-part explanatory hypothesis:[5]

(1) The agent is faced with such-and-such options, which are his/her possible choices.
(2) The agent considers those options, where "considering the options" can take any form ranging from slow and rational deliberation to fast and instinctive processing.
(3) The agent chooses one option among the possible ones, where that choice is intelligible – perhaps rational or approximately rational – in light of the agent's beliefs and goals.

Recall the earlier examples of questions:

- Why, and under what conditions, do people engage in religious practices?
- Why do consumers make the choices they do?
- Why, and under what conditions, do people cooperate with others even when they would benefit from defecting?
- Why do some people vote for populist parties while others do not?

A satisfactory answer to any such question usually fits the above-mentioned three-part explanatory structure (1)–(3). For example, to explain why people engage in religious practices and when, we must ascribe to those people a choice between engaging in religious practices and not doing so. We must also ascribe to them some process of considering their different options and comparing them. And we must point to what makes any person's actual choice intelligible or rational, given the person's beliefs and goals. Similarly, to explain why people cooperate with others under certain conditions, we must ascribe to them a choice between cooperating and not cooperating. We must ascribe to them some process of considering and comparing these options. And we must identify what makes their choice of cooperation intelligible and/or rational.

Any such explanation would not get off the ground if we didn't assume that people make choices between different possible options in the first place. Intentional explanations thus presuppose alternative possibilities. Indeed, intentional explanations are normally *contrastive*: we explain why someone chooses one option *rather than* some others, which would also have been possible for that agent. And we cite something like preferences, reasons, and motivations to render intelligible ("rationalize") the choice in question, in the context of the possible alternatives. The postulate that there is more than one option, among which the agent is making an intelligible choice, is a precondition. The present explanatory scheme can be found, in different versions, across the social sciences.

We find the most paradigmatic version of this explanatory scheme in decision theory, and this is the area to which I will now turn.

[5] Here, again, I draw on my earlier work, especially List (2023).

4. Free will in decision theory

The notion of choice between different options is central for decision theory. Decision-theoretic explanations presuppose that decision-makers face choices and that when they do, more than one option is in principle possible for them, even if only some of the options may be rational. They then seek to explain, predict, or rationalize which options will be chosen. The explanation, prediction, or rationalization typically takes the form of ascribing to the agent some preferences and/or beliefs that would render the choice of some of the options (but not others) rational. The background assumption is:

> **Openness of choices:** An agent's trajectory up to any relevant choice node admits different continuations that are in principle possible for the agent, one of which will be chosen.

This background assumption is formally built into the kinds of extensive-form decision trees modelled in decision and game theory, as illustrated in Figure 1.

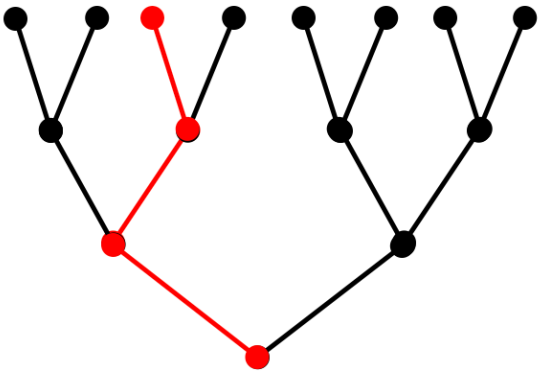

Figure 1: A decision tree

Such a tree represents all the possible decision paths the relevant decision-maker or decision-makers *could* take, moving from bottom to top. The red path – choosing left at the first node, then choosing right at the second node, and subsequently choosing left at the third node – might be the predicted or rational decision path, according to decision theory. However, even if one path is identified as rational, the other paths remain possible. Decision theory draws a distinction between what is possible and what is rational. Without that distinction, the logic of decision-theoretic explanations and/or rationalizations could not be applied.

Any decision tree induces a model of possibilities. One can define the set of "possible worlds", according to a decision tree, as the set of all possible paths through that tree, from bottom to top. For each decision node, one can then define an accessibility relation between worlds (paths) by deeming any world (path) to be *accessible* from another if and only if the two worlds (paths) share the same initial segment up to the given node. Using this accessibility relation, we can define a modal logic of possibility (a so-called S5 modal logic), according to which, at each decision path and node, something is *possible* if and only if it holds at some continuation of the initial segment of the given path up to the given node. (Similarly, something is *necessary*

if and only if it holds at every continuation of the given initial segment.) This modal logic formalizes the idea that at any given decision node the different available continuations of the decision path up to that node are possible.

Indeed, the idea that a criterion of rationality or a game-theoretic solution concept will identify some paths as rational from amongst the possible ones presupposes that there are different such possible paths to begin with. If only one path were possible from the outset, the notion of rationality or any other decision- or game-theoretic solution concept would be trivialized.

Perhaps readers will already be convinced at this point that decision theory presupposes that agents make choices between alternative possibilities and that it thereby presupposes a form of free will. But it is nonetheless useful to look at how denying this presupposition would force us to re-interpret decision theory in an unnatural and non-standard way. I will consider two different uses to which decision theory may be put: normative and descriptive.

4.1 Decision theory as a normative theory

When decision theory is understood as a normative theory, the aim is to determine how a decision-maker (either an individual acting alone or a player in a game) *ought rationally to choose* in certain situations, for instance under uncertainty or in strategic interactions. The assumption is usually that a decision-maker faces a choice between different possible options, such as actions or strategies. We want to know which of these options are rational, under certain conditions, and which not.

If this is meant to offer action-guidance for the decision-maker, i.e., if it is meant to tell the decision-maker how he or she ought rationally to act, the presumption must be that all the options in the decision-maker's *menu* – the set of options available in a given choice situation – are in principle possible. We can then determine, using an appropriate concept of rationality, which option or options the decision-maker ought to choose.

If we assumed that only one option is genuinely possible in any situation, then the decision-theoretic recommendation would

- *either* be trivial, namely in the case of a choice from a singleton menu, which offers only one option,
- *or*, if we still somehow considered a non-singleton "menu", the decision-theoretic analysis would run the risk of violating the "ought implies can" constraint, namely if the rational option is impossible. (The "ought implies can" constraint says that "for any *x*, one ought to do *x only if* it is possible to do *x*".)

Alternatively, the point of decision theory could no longer be

(i) to determine which option, *among several possible ones*, is *rational*,

but merely

(ii) to determine which option, *among several apparent but false possibilities*, is *genuinely possible*, while the others are impossible.

This goes very much against the standard interpretation of normative decision theory, according to which the point of decision theory is (i) rather than (ii). While it is easy to see that the aim expressed by (i) can characterize normative decision theory, it is hard to see how the aim expressed by (ii) could do so.

Moreover, the assumption that only one option is genuinely possible in any choice situation goes totally against the standard decision-theoretic distinction between what is possible and what is rational. Normally, what is rational does not coincide with (but rather corresponds to a subset of) what is possible or feasible.

I conclude, therefore, that normative decision theory, as ordinarily understood, would be in significant trouble if we denied that decision-makers have alternative possibilities to choose from. Let us move on to decision theory in the descriptive or predictive sense.

4.2 Decision theory as a descriptive or predictive theory

Here, the aim is to explain or to predict the actual choice behaviour of some decision-maker or decision-makers, as observed or observable in the real world, such as in the field or in a decision-theoretic laboratory. The typical *explanandum*, i.e., the thing to be explained, or the *target of the prediction*, i.e., the thing to be predicted, is a decision-maker's observed or observable choice function (for an overview, see, e.g., Bossert and Suzumura 2010). A *choice function* is a mapping that assigns to each menu in some domain of possible menus a chosen option from that menu (or a non-empty set of options that are tied for choice). I give an example below (Table 1). We are then looking for a theory that explains or predicts that choice function. In the simplest case, the theory might say that the decision-maker always chooses a most preferred option from the menu relative to some preference order.

So, the *explanans* or *predictor* – the thing that provides the explanation or prediction – is a preference order or some richer construct such as a belief-preference pair or a utility-subjective-probability pair, together with the hypothesis that in each choice situation the decision-maker chooses an option from amongst the set of possible options that he or she prefers (or weakly prefers) to all the other possible options in that situation. (In more complex versions of this, the decision-maker's preference over the options may be based on expectational reasoning and thus dependent on his or her beliefs.) This is clearly an instantiation of the three-part explanatory scheme discussed earlier.

Neither the *explanans* (or predictor) nor the *explanandum* (or target of the prediction) make much sense without the assumption of alternative possibilities. Let us begin with the *explanans* or predictor. A preference order – to focus on the simplest kind of decision-theoretic *explanans* – is normally defined as a binary relation over different possible options or over different possible outcomes. The explanatory or predictive hypothesis, then, is that in any choice

situation, where the decision-maker has a particular menu of options, he or she chooses a most preferred option from that menu. This way of explaining or predicting choices, at least when literally understood, presupposes that the menu consists of options that the decision-maker could in principle choose, i.e., options that are open to him or her. Options that are not possible choices, for instance by failing to meet certain feasibility constraints, are normally excluded from the menu. So, our explanatory or predictive hypothesis is normally understood as follows:

> **H:** Whenever the decision-maker can choose between different possible options, the decision-maker will choose an option that he or she most prefers (according to an underlying preference relation).

For example, we want to explain, by reference to the decision-maker's preferences over the different possibilities, why that decision-maker chooses x when x, y, and z are possible. Or we want to predict which option he or she will choose when x, y, and z are possible. Hypothesis H is both intelligible and familiar. Of course, H is a very simple example of a decision-theoretic hypothesis, which I am using here just for illustrative purposes. Decision-theorists, for example in behavioural economics, often formulate more complex hypotheses. The point I want to emphasize (and which carries over to more complex decision-theoretic hypotheses) is that this kind of hypothesis presupposes that the agent has alternative possibilities to choose from.

Suppose, by contrast, we try to reinterpret our explanatory or predictive hypothesis in a way that denies the existence of alternative possibilities. The hypothesis would then have to be something along the following lines:

> **H':** Whenever it *appears*, albeit falsely, that the decision-maker can choose between different options, the only really possible option will be one (from amongst the set of *falsely assumed* options) that the decision-maker most prefers (according to an underlying preference relation).

In hypothesis H', we would further have to clarify *to whom* it appears (falsely) that the decision-maker has a choice between different options: to the decision-maker him- or herself or to the modeller or some other observer (or all of these)? Moreover, in formulating H', we cannot refer to "the option chosen by the decision-maker", but we must use the cumbersome wording "the only really possible option", because, given the lack of alternative possibilities, there is no real choice here.

It is doubtful whether something like H' is genuinely explanatory. At most, it offers a sort of "error theory" of someone's *apparent* choice behaviour. There is no real choice, according to this reinterpreted hypothesis, and the false appearance of a choice is redescribed as if it was the outcome of a maximization exercise over some options of which only one was ever genuinely possible. A good explanation of an observable phenomenon should render that phenomenon more expected and less surprising, and it should not gratuitously rely on postulating falsehoods or errors. A decision-theoretic explanation of someone's choice behaviour that denies that any real choice is being made does not seem to meet this desideratum.

However, it is not only the decision-theoretic *explanans* or *predictor* that fails to make much sense if we deny the existence of alternative possibilities. Even the *explanandum* itself or the *target of the prediction* – namely a decision-maker's choice function – becomes hard to interpret in that case and also hard to justify as an observable basis for testing decision theory. To see this, consider a simple illustrative choice function, defined over all menus from the set of options {apple, banana, coconut}. The choice function – a mapping from non-empty sets of options to chosen options (or non-empty subsets of options tied for choice) – could be as shown in Table 1.[6]

$C(\{\text{apple, banana, coconut}\}) = \{\text{apple}\}$
$C(\{\text{banana, coconut}\}) = \{\text{banana}\}$
$C(\{\text{apple, banana}\}) = \{\text{apple}\}$
$C(\{\text{apple, coconut}\}) = \{\text{apple}\}$
$C(\{\text{apple}\}) = \{\text{apple}\}$
$C(\{\text{banana}\}) = \{\text{banana}\}$
$C(\{\text{coconut}\}) = \{\text{coconut}\}$

Table 1: A choice function

How should we interpret such a choice function? On the most natural interpretation, it means the following:

- When an apple, a banana, and a coconut are possible, the decision-maker chooses the apple.
- When a banana and a coconut are possible, the decision-maker chooses the banana.
- When an apple and a banana are possible, the decision-maker chooses the apple.
- And so on.

So, the "data points" encoded by the choice function are of the form: when such-and-such options are possible, the decision-maker will choose such-and-such option. A choice function is, in effect, a data set consisting of many such data points, namely one data point for each possible menu of options. Taking such a data set to be the *explanandum* of decision theory or the target of the prediction makes perfect sense on the assumption that the decision-maker has the capacity to choose between alternative possibilities. By contrast, without that assumption, the present *explanandum* breaks away or needs to be reinterpreted.

Recall again the choice function shown in Table 1. How could this function be interpreted if the decision-maker never really has any alternative possibilities for choice? It would still have to be, in some sense, a function that maps "choice situations" to "chosen options", but "choice situations" could no longer be interpreted as genuine *choice* situations, where different possible

[6] Formally, if X is some underlying universal set of options (e.g., $X = \{\text{apple, banana, coconut}\}$), then a *choice function* is a function C which assigns to each non-empty "menu" $Y \subseteq X$ a singleton (or more generally, non-empty) subset of that menu, $C(Y) \subseteq Y$, which consists of the option that is chosen from that menu (or more generally, the options that are tied for choice).

options are open to the decision-maker, and "chosen options" could no longer be interpreted as *chosen* options, since the decision-maker would not have any real choice. Let us run through some non-standard interpretations of our choice function in Table 1 that might be given here.

On one non-standard interpretation, the choice function would encode the following information:

- When *the decision-maker falsely thinks* an apple, a banana, and a coconut are possible, only the apple is really possible (and will be taken).
- When *the decision-maker falsely thinks* a banana and a coconut are possible, only the banana is really possible (and will be taken).
- When *the decision-maker falsely thinks* an apple and a banana are possible, only the apple is really possible (and will be taken).
- And so on.

On this interpretation, no genuine *choice* is being made by the decision-maker. Rather, only the actually taken option will ever have been possible to begin with. For there to be a choice in any non-trivial sense, different options must be possible to begin with. Here, there never really is a non-singleton menu of possible options, only a non-singleton menu of *falsely assumed possibilities*.

According to this re-interpretation, the goal of decision theory, now of the descriptive or predictive sort, would have to be redefined. It would no longer be

(i) to explain and/or predict which option, *among several possible ones*, will be chosen,

but merely

(ii) to explain and/or predict, for each situation, why such-and-such option is the uniquely possible one when the decision-maker falsely thinks such-and-such non-singleton set of options is possible.

Perhaps this re-interpretation could be made coherent, but it is far from the usual interpretation, and far from how decision theory is ordinarily taught. In case of doubt, I challenge readers to introduce students to decision and game theory in a way that avoids the standard interpretation and replaces it with the present non-standard one.

Moreover, the present reinterpretation would make it much less easily observable what a decision-maker's choice function is. If, as is standardly assumed, it is an objective fact which menu of options the decision-maker has in any given choice situation, we can simply present the decision-maker with each menu and observe what choice he or she makes. However, if the options on the menu are not real possibilities but merely options that the decision-maker *falsely believes to be possible*, observing a choice function would require us to know what the decision-maker believes about his or her options. What was previously supposed to be observable by looking at the decision-maker's behaviour will now be observable *only if* we can

come to know the decision-maker's beliefs too. A menu is no longer something objective, but merely something falsely imagined by the decision-maker.

On another non-standard interpretation, the choice function could mean the following:

- When *the modeller (or some other relevant outside observer) falsely thinks* an apple, a banana, and a coconut are possible, only the apple is really possible (and will be taken).
- When *the modeller (or the relevant outside observer) falsely thinks* a banana and a coconut are possible, only the banana is really possible (and will be taken).
- When *the modeller (or the relevant outside observer) falsely thinks* an apple and a banana are possible, only the apple is really possible (and will be taken).
- And so on.

Here, again, no genuine choice is being made by the decision-maker, since only the actually taken option (we cannot say "the actually *chosen* option") will have been possible to begin with. Once more, the goal of decision theory would have to be reinterpreted. It would be

(iii) to explain and/or predict, for each situation, why such-and-such option is the uniquely possible one when the modeller (or other relevant outside observer) falsely thinks such-and-such non-singleton set of options is possible.

Again, this is a very non-standard interpretation. Although the reference to the decision-maker's false beliefs about the options has been removed, it has been replaced by a reference to the modeller's or other observer's false beliefs. Again, this interpretation presupposes a kind of error theory about the *explanandum* of decision theory, which is very different from the natural interpretation.

Matters become even more complicated when – as decision theorists sometimes do – we recognize that there may be discrepancies between the modeller's (objective) model of the choice situation and the decision-maker's (subjective) awareness of it. Decision theorists sometimes study what a decision-maker will do when he or she is unaware of some of the options that are really available. Here, what *seems possible* to the decision-maker differs from what is *really possible* from the modeller's objective perspective. The usual background assumption is that the modeller has better (i.e., more complete and more accurate) information about the choice situation than the decision-maker does. On the no-alternative-possibilities interpretation, however, we would have to reinterpret this scenario in a very odd way. We would have to say that this is a case where the modeller falsely thinks that more options are possible than falsely considered possible by the decision-maker. And since only one option (the actually taken one) would be genuinely possible, we would have to say – oddly – that the decision-maker, with his or her narrower set of falsely assumed possibilities, has a better representation of the true "choice situation" than the modeller. It would not be that the decision-maker is unaware of some possibilities that objectively exist; rather, the modeller would make the mistake of assuming the existence of additional options that are, in reality, impossible.

Furthermore, these non-standard interpretations of a choice function would go against a core methodological tenet of decision theory in economics and other social sciences, namely a methodological tenet that is central to the so-called "revealed-preference" paradigm. Its key idea is that while preferences and other mental states, such as beliefs, are not observable (and certainly not directly observable), choice behaviour *is* observable. Choice behaviour may indirectly "reveal" a decision-maker's preferences and other mental states such as beliefs but only in the sense that some hypotheses about a decision-maker's preferences and/or beliefs are consistent with a particular observable pattern of choice behaviour while other such hypotheses are not. As noted earlier, *if* it is an objective fact which options the decision-maker has in any choice situation, we can in principle elicit and thereby observe a choice function by presenting the decision-maker with each possible menu of options and observing what choice he or she makes. On the assumption that in any choice situation it is sufficiently objective what the decision-maker's options are, choice functions are thus observable. They are then functions that map objective choice menus to chosen options, and that is exactly how economists usually interpret them. However, according to the alternative interpretations I have considered here, menus consist of falsely imagined options, and choice functions then become mappings from belief sets (such as the decision-maker's false beliefs about his or her options) to seemingly chosen options. Choice functions would cease to be purely behavioural and would become mentalistic, which goes very much against the mainstream understanding of choice functions in the revealed-preference paradigm. From the perspective of economics in particular, this may be one of the greatest methodological costs of giving up the presupposition of real alternative possibilities and replacing it with a notion of falsely believed alternative possibilities.

Could one try to restore the observability of a decision-maker's choice function by saying that something counts as an option for a given decision-maker in a given situation if and only if *someone else in that situation (if not the decision-maker him- or herself)* would have chosen it? Although this would give rise to a more complicated interpretation of "option availability", it might seem to restore the observability of a decision-maker's choice function. One could in principle put different decision-makers into the same choice situation and observe what choices they make. An option would count as being "available in the given situation" if it is chosen by at least one such decision-maker.

On closer inspection, however, this interpretation is also problematic. Its core difficulty is that, unless we have a prior understanding of what it means for a given set of options to be available, we lose a good grip on when two distinct decision-makers are in fact in the same choice situation. On the assumption that only the actually taken option is genuinely possible, it is no longer clear how two distinct decision-makers could count as being "in the same choice situation" if they end up taking (we can't say "choosing") different options, which are their only real possibilities.

To see this, first consider the standard assumption that each decision-maker really has all the options on the menu that is presented to him or her. On this assumption, it is perfectly intelligible how Alice and Bob could be in the same choice situation even though Alice ends up choosing an apple while Bob chooses a banana. To count as being in the same choice

situation, they would simply have to have the same menu of options. On the standard assumption, both options would indeed be open to both decision-makers, and it is then in light of their different preferences that they make different choices.

However, if Alice's only real possibility is the apple while Bob's only real possibility is the banana, it is harder to see in what sense they could be *in the same choice situation*. The criterion for being in the same choice situation could no longer be "having the same possible options", since Alice and Bob wouldn't have the same possible options here. At most, their *falsely imagined* possibilities would be the same, but we wanted to get away from defining choice situations in terms of a decision-maker's false beliefs about the options.

Suppose, on the other hand, "being in the same situation" means "having the same menu of *available* options", where the menu of *available* options is re-defined technically in terms of what *either oneself or another decision-maker* would choose in that situation. We then end up with a circular definition. To illustrate this, suppose we say that Alice, Bob, and Chiara are in the same situation because they each have the same menu of available options, and we say that they have the same menu of available options because they are each in a situation in which

- someone (namely, Alice) would choose an apple;
- someone (namely, Bob) would choose a banana; and
- someone (namely, Chiara) would choose a coconut.

Given that they make different choices and each decision-maker's choice is the only possibility that he or she really has, it is hard to explain how they could all be in the same choice situation unless we sneak in the background assumption that the same three options are still somehow "available" to all of them. But we are not allowed to make this background assumption, since we are only trying to *define* what it means to say that these three options are "available" to them. In short, it is hard to define the availability of certain options in terms of a thought experiment about what other decision-makers would do in the same situation unless we already presuppose that "sameness of the situation" entails "sameness of the available options".

Here is another way in which one might try to develop the view that something counts as an option for a given decision-maker in a given situation if and only if some (possibly other) decision-maker in that situation would have chosen it. One might hypothetically imagine a perfect "duplicate" of the world in which the given decision-maker, say Alice, is located, but with the important modification that Alice has been replaced by Bob, while everything else remains equal. If, in that alternative possible world, Bob chooses/takes a banana, we conclude that the banana is also available to Alice in the original world.

As a philosophical thought experiment, something along these lines might make sense. But it yields an interpretation of a choice function that makes choice functions practically impossible to observe. Indeed, the duplicated worlds to which the thought experiment refers may be purely hypothetical. Suppose Alice faces the choice of whether to marry Bob or not. Is it genuinely

realistic (or even fully coherent) to imagine a duplicated world that is perfectly identical to the actual world except that Alice has been replaced by (say) Chiara?

I conclude, therefore, that it is also unsatisfactory to define the availability of some option for a decision-maker in a given situation by reference to what some (possibly other) decision-maker would have chosen in that situation. In sum, giving up the assumption of alternative possibilities, as we conventionally understand it, seriously challenges both the *explanans* and the *explanandum* of decision theory.

5. From explanatory indispensability to realism about free will

I have argued that decision theory presupposes free will. Decision-theoretic explanations depend on the postulate that decision-makers face choices and that when they do, more than one option is in principle possible for them, even if only some of the options may ultimately be rational and/or chosen. Decision theory, like other theories that offer intentional explanations of human behaviour, would be in trouble without that postulate.

But what follows from the explanatory indispensability of the free-will postulate? Perhaps this postulate is merely a theoretical construct or an instrumentally useful fiction. Decision-theoretic explanations, on this interpretation, should be viewed as giving us "as if" stories that may be intuitively intelligible but that shouldn't be treated as literally true. Gilboa (2007) seems to have something like this in mind when he describes free will as an "illusion" that is "essential to rational decision making" (p. 1). Indeed, economists often interpret decision-theoretic explanations in this "as if" mode: decision-makers behave *as if* they maximize utility, for example, even if they aren't truly utility-maximizers.

However, an instrumentalist, fictionalist, or illusionist interpretation of the free-will postulate is out of line with how explanatorily indispensable postulates are treated elsewhere in science. In the natural sciences, it is common to treat explanatory indispensability as an indicator of reality. That is: if our best explanations of some phenomena depend on postulating certain properties or entities, we tend to conclude that those properties or entities are real. Why, for example, do we think that electromagnetism and gravity are real? We cannot *directly* observe either electromagnetism or gravity. Rather, there are some *other* observable phenomena that are best explained if we *postulate* electromagnetism and gravity: for example, how macroscopic objects fall when dropped or how a bicycle lamp lights up when a dynamo is operated. We regard the explanatory indispensability of postulating electromagnetism and gravity as an indicator that those postulates correctly describe the world. Similarly, when our best scientific explanations of why certain particles have mass require us to postulate the Higgs boson, physicists tentatively conclude that the Higgs boson must exist. Of course, this kind of reasoning is always provisional. New discoveries may lead us to revise our explanations and prompt us to revise the postulates in question. In the meantime, however, scientists tend to consider the postulates of our best scientific theories to be at least approximately correct.

To be consistent in our scientific methodology, we should interpret the postulates of our best theories in the human and social sciences in the same way.[7] For example, if our best explanations of human behaviour postulate that people have beliefs and preferences, we have every reason to think that people really have beliefs and preferences. Of course, our best explanations may still make simplifications, and people's beliefs and preferences may be more complex in the real world than in our models. Nonetheless, it would be surprising if the postulate that people have beliefs and preferences were so useful in our explanations if it was a fiction, just as it would be surprising if the postulate that there is gravity were so useful if gravity was a fiction. As Hilary Putnam (1975, p. 73) argued, a realist stance towards the sciences is "the only philosophy that doesn't make the success of science a miracle". Wouldn't it seem miraculous, for instance, if physical objects behaved *as if* they were subject to gravity if gravity wasn't real?

Returning to the case of human decision-making, one may similarly argue that *if* the postulate that humans have free will – understood as the conjunction of intentional agency, alternative possibilities, and causal control – is explanatorily indispensable in the sciences of human behaviour, this gives us good reasons for taking this postulate to be approximately correct. How could postulating the human capacity to make free choices be so explanatorily useful if humans lacked that capacity?

Some readers may still object to the analogy with physical phenomena such as gravity, electromagnetism, or the Higgs boson. The existence of all these phenomena, they may say, is not just a postulate but has been observationally confirmed. However, what has been observationally confirmed is not gravity itself, or electromagnetism itself, or the Higgs boson itself. None of these are directly observable. Rather, what has been observationally confirmed in each case is a broader *theory* that relies on postulating the given property or entity. If we postulate gravity, we can adequately explain many observable phenomena, such as how macroscopic objects behave under various conditions. Without postulating gravity, we lack a good explanation. Likewise, if we postulate electromagnetism, we have a good explanation of the phenomena addressed by Maxwell's theory of electrodynamics. Without postulating electromagnetism, we lack a good explanation. And with the postulate of the Higgs boson, we can explain why certain particles have mass; without it, we lack that explanation.[8]

---

[7] In List (2014, 2019), I argued for this point by suggesting that we should adopt the "naturalistic ontological attitude" (Quine 1977, Fine 1984), i.e., take the ontological commitments of our best scientific theories in any given domain as a guide to which entities or properties are real.

[8] The objector might still insist that the Higgs boson has been experimentally confirmed. Wikipedia, for instance, says: "After a 40-year search, a subatomic particle with the expected properties was discovered in 2012 by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) at CERN near Geneva, Switzerland" (https://en.wikipedia.org/wiki/Higgs_boson, accessed 19 July 2024). However, this is an informal gloss that omits the fact that the Higgs boson is not visible itself, even with the help of the very best microscopes, and that our evidence for it is only indirect. Certain large data sets from complex experiments in particle accelerators such as the Large Hadron Collider are *best explained* by a theory that postulates the existence of the Higgs boson. Among other things, this is because the Higgs boson is very unstable and decays very quickly, so that we can only detect it by observing its decay products. The data collected by CERN suggests that a particular decay process took place

The case of gravity may be particularly illuminating. Our best accounts of gravity say something along the following lines: "objects with mass curve the space around them". But we have no *intuitive* idea what this means. Any formalization of the notion of space-time curvature is highly technical and could easily be dismissed as a merely instrumentally useful construct, on the grounds that it may sound counterintuitive and mysterious. Yet, the great explanatory usefulness of postulating gravity leads most scientists to think of gravity as real.

In the same way, I argue, the fact that human behaviour – across a variety of domains – is best explained by theories that presuppose that people are intentional agents, with alternative possibilities to choose from, and relevant control over their actions lends support to the hypothesis that people really are agents with free will. To be sure, the step from complex neural firing patterns in the brain to free agency may seem intuitively hard to grasp, just as the mechanisms underlying gravity – spelt out in terms of space-time curvature – may seem intuitively hard to grasp. Nevertheless, just as in the one case we treat explanatory indispensability as evidence of reality, we should do so in the other. The support for free will, like the support for gravity, is indirect and comes from an inference to the best explanation.

One way to avoid this kind of realism about free will would be to endorse a thoroughly instrumentalist view about all the sciences and to take an anti-realist stance towards *every* scientific postulate that is not directly observable. One would then have to say that gravity, electromagnetism, and the Higgs boson are merely instrumentally useful constructs that do not genuinely exist, and one would have to say the same about all other properties and entities for which we have only indirect support, no matter how well-confirmed the relevant theories are. A proponent of that view could reject the reality of free will too, despite its explanatory indispensability in the sciences of human decision-making. But instrumentalism is neither the dominant view in the philosophy of science, nor (arguably) the most defensible one.

In any case, while thorough-going realism in science and thorough-going instrumentalism are each internally consistent views, it would not be very principled to be a realist about gravity, electromagnetism, and the Higgs boson, and perhaps about organisms, economies, and income distributions and yet to be an anti-realist about human agency and free will. It is not consistent to cherry-pick one's philosophical attitudes here, by being a realist about some kinds of indispensable postulates, such as gravity and electromagnetism, while being an anti-realist about others, such as agency and free will.

6. Free will is not unpredictability

I have argued that decision theory presupposes free will, especially the ability to make choices between alternative possibilities, and that there are good scientific reasons for assuming that this postulate is true. According to the most natural interpretation of decision theory, any decision-maker, as modelled by decision theory, has a free choice at every choice node that he

---

that matches the predicted "decay signature" of the Higgs boson, according to the standard model of particle physics. Scientists' belief in the existence of the Higgs boson thus relies on an inference to the best explanation.

or she faces. That is: there is a "fork in the road" ahead of the decision-maker at each choice node. In Figure 1, for example, it is possible for the decision-maker to go either left or right at each node, even if it turns out that only one path is ultimately rational for the decision-maker. However, this "libertarian" picture of choice-making invites a significant objection. Doesn't the claim that a decision-maker has free will imply that the decision-maker is unpredictable? And conversely, wouldn't the fact that decision-makers are often predictable undermine the claim that they have free will? Even more importantly, if the aspiration of the human and social sciences is to explain and predict human behaviour, wouldn't this mean that the more successfully our theories live up to this aspiration, the less they will be compatible with the hypothesis that human beings have free will?

Both Itzhak Gilboa and Scott Aaronson make versions of this point. Both suggest that if a decision-maker or an outside observer could know or predict what choices the decision-maker will make, this would imply that, even if the decision-maker *contemplates* alternative possibilities, the choice is not in fact free.

Gilboa (2007, pp. 4–5), for instance, writes:

> "We know decisions that we have made, and we can often have pretty good guesses about certain decisions that we are going to make. I know that I'm going to prefer coffee to tea. I know that I prefer not jumping out of the window to jumping. As a rational decision maker, I gather data and make inferences. I cannot help observe regularities around me, and my own decisions in the past are included in the environment I study. Moreover, it is essential for rational choice that I learn things about myself. I need to know my 'technical' capabilities, such as how fast I can run and how good my eyesight is. It will also be useful to know something about my mental capacities, such as how good my memory is and to what extent I follow my new year's resolutions. For this latter purpose, I need to know my own choices in circumstances in which I felt that I was exercising free will. Finally, learning regularities about myself can be useful in predicting other people's behavior. …
>
> [R]ationality makes two fundamental demands. First, we have to consider possible worlds that differ in terms of our choices. Second, we have to observe obvious regularities about ourselves, just like about any other relevant phenomenon. Taken together, we obtain the contradiction: we often need to consider as possible worlds that we know are impossible."

Similarly, Aaronson (2013) suggests that the "central question" concerning human freedom is "how well complicated biological systems like human brains can actually be predicted: not by hypothetical Laplace demons, but by prediction devices compatible with the laws of physics" (p. 5). He writes:

> "I'll use the term freedom, or Knightian freedom, to mean a certain strong kind of physical unpredictability: a lack of determination, even probabilistic determination, by

> knowable external factors. That is, a physical system will be 'free' if and only if it's unpredictable in a sufficiently strong sense, and 'freedom' will simply be that property possessed by free systems. A system that's not 'free' will be called 'mechanistic.'" (p. 7)

Although he takes no view on whether there is freedom in this sense, he thinks that

> "if it turned out … that human brains were as probabilistically predictable by external agents as ordinary digital computers equipped with random-number generators[,] then [the hypothesis that there is human freedom in this sense] would be falsified, to whatever extent it says anything interesting." (p. 6)

According to this line of reasoning, a theory that allows us to make good predictions of human behaviour would immediately speak against the hypothesis that human beings have free will. Suppose we concede, consistently with my argument, that the human and social sciences, and especially decision theory, are often able to make reasonably accurate predictions of human behaviour, and that they are getting better as science progresses. Should this not lead us to abandon the claim that human beings have free will? And would it not illustrate the sort of deep inconsistency that, according to Gilboa, lies at the heart of decision theory?

I think this reasoning is mistaken. Free will does not imply unpredictability. Unpredictability is neither necessary nor sufficient for free will. It should be obvious that it is insufficient. Unpredictability might stem from randomness, and randomness does not entail free will. The weather or solar flares are, to a good extent, unpredictable, but weather systems or the sun do not have free will. Stochastic explanations of those phenomena are explanatorily superior to intentional ones.

But unpredictability is also not necessary for free will. Let me first give an example to support this point before clarifying my position further. It so happens that I do not drink alcohol. Anyone who knows me well knows this. So, it is predictable that whenever I go into a bar or restaurant, I will not choose an alcoholic beverage. Suppose that I have a choice between two drinks: an alcoholic drink and a non-alcoholic one. My choice is completely predictable here. Note further that we can predict my choice not using neuroscience but by reference to my preferences. Yet, this does not undermine my free will. When I face a choice between different such drinks, it remains a choice: I have intentional agency, both options are open to me, and I have control over the resulting action. Although I reliably choose the non-alcoholic beverage, I could choose otherwise. The choice is mine.

Free will is thus fully compatible with rational predictability of an agent's choices. We must not conflate the possibility of acting otherwise, which is required for free will, with unpredictability or randomness, which are not required. Rational predictability of my choice does not imply a lack of alternative possibilities. It is possible for me to act otherwise and thereby to frustrate the prediction, even if I actually do not do so. Of course, if I were to choose the alcoholic beverage, I would act against my own preferences, and so I would be irrational by my own lights. But irrational choices are not impossible. The same is true in decision and

game theory. Even if we identify a particular series of choices as uniquely rational in a decision- or game-theoretic model of a particular choice problem, this does not mean that the relevant decision-makers should be viewed as lacking free will.[9] From a decision- or game-theoretic perspective, it is still possible for them to act otherwise. Off-equilibrium paths in extensive-form decision trees are not assumed to be impossible, merely irrational. So, predictability can go along with the existence of alternative possibilities. Note, further, that unpredictability would not generally be helpful for free will. If my choice were random, for instance, it would be unpredictable but hardly attributable to my own free will.

It is important to emphasize that my argument for free will is *not* the following:

> **Premise 1:** Human behaviour is unpredictable in a sufficiently strong sense.
>
> **Premise 2:** If human behaviour is unpredictable in a sufficiently strong sense, there is free will (or at least, a key necessary condition for free will is met).
>
> **Conclusion:** There is free will (or a key necessary condition for free will is met).

Rather, my argument is broadly the following:

> **Premise 1:** Human behaviour can be reasonably well explained and/or predicted.
>
> **Premise 2:** Human behaviour can be reasonably well explained and/or predicted only if we understand human beings as choice-making agents.
>
> **Premise 3:** If we understand human beings as choice-making agents, we are relying on the postulate that human beings have free will.
>
> **Premise 4:** If some postulate is explanatorily indispensable (i.e., certain phenomena can be reasonably well explained and/or predicted, but this is so only if we are relying on that postulate), then we have provisional reasons for taking the postulate to be true.
>
> **Conclusion:** We have provisional reasons for taking the postulate that human beings have free will to be true.

Therefore, far from requiring unpredictability, on my analysis, free will is actually a presupposition of our best approaches to explaining and predicting human behaviour.

To reinforce this point, it may help to distinguish between two different forms of prediction. To do so, recall the distinction between dynamical and intentional explanations. In a dynamical explanation, as in physics, we explain a system's behaviour by describing deterministic or stochastic state-change rules by which a system changes its state over time, as for instance in a Newtonian model of the solar system. These state-change rules encode the laws of nature

[9] Note also that there may be multiple equilibria anyway, and so even the rational path need not be unique.

governing the system in question. In an intentional explanation, by contrast, we take what Dennett (1987) calls an "intentional stance" towards the object of our explanation, which for present purposes is human behaviour (though biologists also sometimes explain non-human behaviour in this way; see, e.g., Tomasello 2022). Intentional explanations explain human behaviour by understanding humans as intentional agents capable of making at least approximately intelligible choices. Intentional explanations, as I have argued, presuppose a form of free will, understood as the conjunction of intentional agency, alternative possibilities to choose from, and causal control over the resulting actions. Such explanations won't get off the ground unless we presuppose that the agents in question can make choices.

Now contrast two corresponding forms of prediction. Suppose a system admits a dynamical explanation. If we know the system's state at a particular time, we may be able to plug that information into our dynamical model of the system and thereby predict its future states. When we use physical models of the solar system to predict the date and time of the next solar eclipse in New York, for example, this is essentially what we do. Call such predictions *nomological predictions*. They are predictions about how a system will evolve under the laws governing it. Nothing like choice ever comes up here; at most, there could be randomness, in case our model is stochastic. A nomological prediction asserts that, given our dynamical model and assuming we have made no mistake, it is *nomologically necessary* for the system to behave in the predicted way, i.e., necessary relative to the laws governing that system. This is essentially how we interpret our prediction of the next solar eclipse in New York. Nomological predictions are based on the idea that the system behaves deterministically or stochastically under the relevant laws, and there is nothing like agency, intentionality, or choice involved. By contrast, if human behaviour – in some domains – is best explained in intentional terms, our explanations presuppose that the people involved are choice-making agents, as I have argued at length. How do we then predict that next time I go into a bar, I will order a non-alcoholic drink? How do we predict that if you agree to meet me for coffee at a particular time next week, you will show up, assuming nothing unforeseen happens (illness etc.)? Or how do we predict that a consumer will choose one consumption bundle rather than another? We do so, not using dynamical models as in physics, but – if we make the logic of the prediction explicit – by ascribing to the relevant agent (or agents) a decision tree as shown in Figure 1 and suggesting that a particular path through that tree will be rational or at least boundedly rational, applying some decision- or game-theoretic equilibrium concept. The prediction asserts that the predicted decision path should be expected because it is rational or intelligible in the relevant sense, while other paths are not. As already noted, the prediction does not say that the alternative paths are impossible. Therefore, the prediction does not depict the predicted events as *necessitated under the laws of nature*; it depicts the predicted behaviour merely as *rationalized* or *rendered intelligible*. Call such predictions *rational predictions*. Rational predictions leave free will and the possibility of acting otherwise intact. This is not to say that rational predictions need to be worse than

nomological predictions. Some of them can be extraordinarily reliable. Crucially, however, they rest on a presupposition of free will.[10]

In sum, predictability of human behaviour would challenge free will only if intentional explanations of human behaviour were dispensable in favour of dynamical ones and if we then had nomological (as opposed to rational) predictability. But the human and social sciences that best explain and/or predict human behaviour in many domains offer intentional explanations and yield rational predictability. And the latter is not just consistent with free will; it actually presupposes it.

7. Concluding remarks

I have argued that, from a decision-theoretic perspective, it is a mistake to consider free will an illusion. Although decision theorists seldom acknowledge this, decision theory crucially relies on the hypothesis that decision-makers have free will. But rather than worrying that this hypothesis is either false or unscientific, I have suggested, we should embrace it. Realism about free will is consistent with realism about other properties and entities postulated in the sciences.

Decision theorists working in economics and other social sciences might still wonder why they should care about all this. Economists hardly ever mention free will in their scholarly work, for example, and they may be content to construct their decision- and game-theoretic models without worrying too much about their philosophical foundations. They might follow the view expressed by the famous phrase "Shut up and calculate", which is often attributed to the physicist Richard Feynman, but which seems to have been coined by another leading physicist, David Mermin, as an informal gloss of the Copenhagen interpretation of quantum mechanics, apparently without endorsement (Baggott 2021). However, just as this anti-philosophical or a-philosophical attitude is arguably a mistake in physics, so I think it is a mistake in economics and the social sciences too.

First, it would be intellectually unsatisfactory to sweep significant philosophical background assumptions of our scientific explanations under the carpet. If we need to presuppose free will to explain human behaviour, we should be upfront about this. Secondly, we could easily get confused about the status of decision- or game-theoretic predictions if we missed the distinction between nomological predictions (as in the physical sciences) and rational predictions (as in decision- and game-theory). As explained, the logic of the latter is fundamentally different from that of the former, insofar as rational predictions are premised on the idea of free choice and the possibility that decision-makers could choose otherwise. And thirdly, the concept of free will remains hugely important from a normative perspective. This shows up not just in morality and the law, but also in economics and the social sciences. Proponents of consumer sovereignty and libertarian paternalism, for instance, are eager to respect choices that economic agents make out of their own free will. Sugden (2004) refers to "free will" when he discusses

[10] My argument that identifying some options as "rational" or "intentionally endorsed" (and thereby as "predicted") is compatible with alternative possibilities builds on work with Wlodek Rabinowicz (List and Rabinowicz 2014).

criteria for an outcome to be "attributable to decisions that [an agent] has taken knowingly of his own free will", adding that "as a responsible agent, he must acknowledge those decisions as his own" (p. 1019). Similarly, ideas of free will underlie the debate about whether nudging people does or does not respect their freedom (Thaler and Sunstein 2008). And finally, welfare economists have been developing responsibility-sensitive approaches to assessing income and wealth distributions, where inequalities that are attributable to people's free choices are considered less problematic than inequalities that must be attributed to bad brute luck. Fleurbaey (1995), for instance, discusses the thesis that "individuals must live with the consequences of their responsible acts" and acknowledges that the "ethical appeal of [this thesis] may not be completely independent of the concept of free will that is employed". And Roemer (1987) mentions "free will" as one of "four reasons why we may deem a person to be responsible for an action he has done or a trait he has" (p. 240), along with other reasons such as incentive compatibility, moral hazard, and the value of choices (pp. 240–241). The idea, which Roemer then scrutinizes, is that, in some cases, "[a] person is responsible for an action he has taken or a trait that he has acquired because he could have reasonably been expected to have behaved differently" (ibid.). In all those areas, it matters greatly whether people genuinely have free will or whether this idea is an illusion that goes back to a prescientific age.

Evidently, then, economists and social scientists should care about free will, and it should be a welcome conclusion that the idea of free will is well supported by the kind of decision theory that is central to their fields.

References

Aaronson, S. (2013). "The Ghost in the Quantum Turing Machine." arXiv:1306.0159v2.

Baggott, J. (2021). "Calculate but don't shut up." *Aeon*, 6 December 2021, <https://aeon.co/essays/shut-up-and-calculate-does-a-disservice-to-quantum-mechanics>.

Bossert, W., and K. Suzumura (2010). *Consistency, Choice, and Rationality*. Cambridge, MA: Harvard University Press.

Churchland, P. M. (1981). "Eliminative materialism and the propositional attitudes." *Journal of Philosophy 78*(2): 67–90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. Cambridge, MA: MIT Press.

Coyne, J. (2014). "What scientific idea is ready for retirement?" Edge.org.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Eagleman, D. (2015). *The Brain: The Story of You*. Edinburgh: Canongate.

Fine, A. (1984). "The Natural Ontological Attitude." In J. Leplin (ed.), *Scientific Realism*, 83–107. Berkeley: University of California Press.

Fischer, J.M., R. Kane, D. Pereboom, and M. Vargas (2007). *Four Views on Free Will*. Oxford: Blackwell.

Fleurbaey, M. (1995). "Equal Opportunity or Equal Social Outcome?" *Economics and Philosophy* 11(1): 25–55.

Gazzaniga, M. (2011). *Who's in Charge: Free Will and the Science of the Brain*. New York: Harper Collins.

Gilboa, I. (2007). “Free Will: A Rational Illusion.” Working paper, Tel Aviv University.

Harris, S. (2012). *Free Will*. New York: Simon and Schuster.

Kane, R. (2011). *The Oxford Handbook of Free Will* (2nd edn). Oxford: Oxford University Press.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Libet, B., C. A. Gleason, E. W. Wright, and D. K. Pearl (1983). “Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act.” *Brain* 106: 623–642.

List, C. (2014). “Free Will, Determinism, and the Possibility of Doing Otherwise.” *Noûs* 48(1): 156–178.

List, C. (2019). *Why Free Will is Real.* Cambridge, MA: Harvard University Press.

List, C. (2023). “Agential Possibilities.” *Possibility Studies and Society* 1(4): 461-470.

List, C., and P. Menzies (2009). “Non-reductive Physicalism and the Limits of the Exclusion Principle.” *Journal of Philosophy* 106(9): 475–502.

List, C., and W. Rabinowicz (2014). “Two Intuitions about Free Will: Alternative Possibilities and Intentional Endorsement.” *Philosophical Perspectives* 28: 155–172.

Pereboom, D. (2001). *Living without Free Will*. Cambridge: Cambridge University Press.

Putnam, H. (1975). *Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Quine, W. V. (1977). *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Roemer, J. E. (1987). “Egalitarianism, Responsibility, and Information.” *Economics and Philosophy* 3(2): 215–244.

Sapolsky, R. (2023). *Determined: A Science of Life without Free Will*. London: Penguin.

Sugden, R. (2004). “The Opportunity Criterion: Consumer Sovereignty without the Assumption of Coherent Preferences.” *American Economic Review* 94(4): 1014–1033.

Thaler, R. H., and C. R. Sunstein (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. London: Penguin.

Tomasello, M. (2023). *The Evolution of Agency*. Cambridge, MA: MIT Press.

Van Inwagen, P. (1975). “The Incompatibility of Free Will and Determinism.” *Philosophical Studies* 27(3): 185–199.