Free Will: A Rational Illusion *

Itzhak Gilboa†

October 2007

Abstract

It is argued that free will is a highly problematic concept, even if no form of determinism is assumed. Free will is an illusion, requiring that one would suspend knowledge about oneself. This illusion is, however, essential to rational decision making and can be justified from an evolutionary viewpoint.

1 Introduction

Discussions of free will often focus on its conflict with determinism. Some recent, scientifically-informed contributions accept the view that determinism is the main challenge to the existence of free will. Penrose (1997) argues that Heisenberg's Uncertainty Principle may suffice to evolve into uncertainty about people's decisions, thus salvaging the notion of free will. Searle (2004) claims that our understanding of neurobiology at present does not yet prove that the brain is deterministic and that free will is an illusion, though he speculates that neurobiological research will get to this point.

This note takes a decision-theoretic approach and claims that the problem of free will is much more pervasive than these accounts suggest. Specifically, even if no form of determinism is assumed, free will is often an illusion. At

*The content of this note is probably not new. I will be grateful for references. I am grateful to Ilana Arbel and Arik Roginsky for comments.

†Tel-Aviv University, HEC, and Cowles Foundation, Yale University. igilboa@tau.ac.il

the same time, the same approach explains why the illusion of free will is necessary to any notion of rational choice, and, perhaps, why we evolved to have this illusion.

2 The Problem

To show the difficulty with free will, one need not assume that *all* decisions are *pre-determined*. It suffices that *one* decision be *known*. The logic is similar to suggesting a counter-example to a conjecture. The existence of one counter-example suffices. Similarly, if we can find one instance in which we have an undeniable sensation of free will on the one hand, and practically certain knowledge of our choice on the other, we will have to admit that the sense of free will is illusionary, at least in this example. In principle, one such example would suffice to put the notion of free will in doubt. In practice, we maintain that such examples abound.

Consider the following example. Sir Isaac Newton stands by a large window on the fourth floor. He contemplates the possibility of jumping out of the window. Should he jump, he considers two possibilities: he may hover in the air, enjoying the view, or crash to the ground. Being a rational decision maker, Newton contemplates the possibility of jumping and, given his knowledge of physics, concludes that crashing to the ground is a practical certainty. He now considers his own decision, and decides not to jump. In so doing, he feels that he has made a decision, and that he has exercised his free will. He could imagine choosing differently, and decided not to.

Suppose that we are sitting with Sir Newton in his office throughout this process. Our limited knowledge of physics suffices for us to conclude, as does Newton, that a jump will result in a crash. With a lesser degree of certainty, but still quite confidently, we are willing to predict that Newton will not jump. We have seen many people next to many windows, and, for the most part, they prefer to stay in their rooms. In short, we know Newton's choice

with a high degree of certainty.[1]

But what about Sir Isaac Newton himself? Surely he knows himself at least as well as we know him. If we could conclude, based on our knowledge of human nature in general, that Newton will not jump, so can he. In fact, he is even in a privileged position to make predictions about himself.[2] Let us examine his reasoning process. A reasoned decision is supposed to take into account rules and regularities that are known to be quite accurate, to help us think about the consequences of our choices. We could imagine Newton drawing a decision tree, and using all his knowledge to assign probabilities to the various branches in the tree, and, in particular, to cross out branches that he knows are practically impossible. This is how Newton concluded that, due to the gravitational force, he will not hover in the air should he jump. But, by the same logic, Newton can now cross out the branch "I jump" just as he previously crossed out the branch "I hover in the air" (conditional on jumping). By the time he finished the analysis there is no longer any decision to be made. Newton knows what his decision will be in the same sense that he knows what the choices of different decisions would be. When was a decision *taken* in this process? And how can Newton report an experience of free will if he cannot imagine a logically consistent world in which he chooses differently? How can we make sense of his claim "but I *could* have jumped"?

[1] One may prefer to use the term "belief" in this context. The point is that this is a high degree of belief, which is probably as high as we can hope for in the social sciences, and higher than our belief in, say, the weather forecast for the day after tomorrow. I do not think that the notion of free will can hinge on events that are possible but improbable, such as zero probability events. One argument against a zero-probability event is aesthetic. It seems cheap. The other is more pragmatic: a zero probability event will not be worth contemplating for even a negligible amount of time. The rational and evolutionary arguments below can be re-stated when "knowledge" is replaced by "belief with very high probability".

[2] Some people have suicidal tendencies, but the majority do not. Our knowledge about Newton, based on statistics on a larger population, is less accurate than his own. Thus, for the majority of individuals it is true that they know that they are not suicidal with a higher degree of certainty than an outside observer would. Since we seek an example, we are justified in assuming that Newton is in this majority.

The paradoxical nature of free will stems from the co-occurrence of (i) the ability to imagine possible worlds that differ in terms of our choices, and (ii) the fact that often our choices is practically known to us before we make it. Let us elaborate on these.

(i) Whatever free will is, it is tightly related to the ability to conceive of different possible worlds, differing only in one's choice and its consequences. The ability to think of such different worlds, if not simultaneously then at least in the process of making a single decision, is essential to rational choice. And this ability is essential to, and maybe even a definition of, the sensation of free will. I feel that I exercise free will when I raise my arm, but not when my heart beats. The reason is that, when consciously deciding to raise my arm I can simultaneously imagine two worlds, one in which the arm is raised and the other in which it isn't. By contrast, I have never felt my heart stopping to beat, let alone decided to do so, and I cannot imagine a choice that would lead to this state of affairs. I therefore cannot argue that I exercised free will in letting my heart beat.

To see this point more clearly, suppose that you program a robot that will automatically make all the choices I make. Next you allow the robot to speak, and you want it to utter the statement "I hereby exercise free will" at the right moments, say, when I make such statements. Let us be slightly more demanding and require that the robot print out reasonable justifications for its choices. To this end, you will have to endow the robot with some reasoning ability, and with the ability to distinguish between its own acts and the environment it lives in. When facing an act, the robot will have to play around with some propositions of the form "If I do a, then the outcome will be x", and "conclude" that it prefers act a to b. The robot will have to print several different such conditional statements for us to agree that it has exercised free will.

(ii) We typically know many things about ourselves. 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.

Let us consider the robot again. Will it know its own choices? Since you are programming it, you may try to avoid such knowledge. It is possible to restrict the inferences made by the robot to external events, and to abort any calculation that refers to the robot's own choices. This will be somewhat artificial. Moreover, it will be inefficient, because the robot will not be able to use its own past decisions as guidance. Every time it will be offered coffee or tea it will have to make a calculation afresh. But the main difficulty with such a robot will be that it will not be as rational as I am. There will be some obvious inferences that it will fail to draw. Our own reasoning engines do not stop when it comes to our own choices in the past. We do learn about ourselves, and someone who fails to see obvious regularities in her own behavior is typically viewed as irrational.

We conclude that rationality 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. Thus, the sensation of free will depends on our ability to suspend knowledge that

we have about ourselves. Importantly, both the consideration of multiple possible worlds and the knowledge that some of them are impossible are dictated by rationality.

3 A rational illusion

At the risk of belaboring obvious points, let me emphasize the following. Not every decision will be known to the decision maker or to an outside observer before it has been taken. As long as the decision maker does not know what her choice is going to be, her sense of free will does not require that she suspend any knowledge she might have. In such a case the problem mentioned above does not exist.

For example, assume that I have to choose between two quantities of a desirable good. We may think of tens of thousands of dollars, or of years left to live – the point is that I truly prefer more to less. Consider now the following three choices:

(i) $\sqrt{17}$ or $(27/13)^2$

(ii) 2^3 or 3^2

(iii) 0 or 1.

In case (i) there is no difficulty. Reading the problem, it is not obvious to me which of the two numbers is larger. I therefore have to compute the outcome of both of my choices, and then find out which one I prefer. An outside observer may have completed the calculation earlier, and may already know what my choice will be. But I do not, and therefore my sense of free will does not contradict any knowledge I have at the time of starting my deliberation.

By contrast, case (iii) is one in which I know, more or less as soon as I read the problem, what my choice will be. I don't need a lengthy computation to figure out the meaning of 0 and of 1. This is akin to Newton's problem, who stands by the window and has to decide whether to jump out or not.

(The analogy is stronger if the numbers designate years one has to live, and 0 describes immediate death.) In both cases one needs to understand the two options and what they entail, but this understanding is quite trivial. The calculation that $1 > 0$ is about as immediate as the realization that jumping out of the window would result in death.

Case (ii) is brought as an intermediate case, suggesting that we cannot think of cases (i) and (iii) as qualitatively different. There is a range of difficulty levels, and a reasonable account of rational choice should describe a process that applies in all three cases. Thus, in all three cases we would like to assume that the decision maker makes a tentative assumption that she takes one option, and thinks about the outcome. Then she does the same for the other option(s), and then she can make a reasoned decision. Whereas in case (i) there is no conflict with knowledge of her own choices, in case (iii) there is. Thus, in cases such as (i) the decision maker may believe that she has free will, but in cases such as (iii) she has to admit that this was an illusion.

Efficiency of decision making might suggest that we need not compute our optimal choice every time anew. We may develop habits and rules that simplify our lives. It would therefore be tempting to categorize all decisions into two classes – the habitual decisions, such as in case (iii), in which there is no freedom of choice, but also no subjective sensation of free will, and the reasoned decisions, such as case (i), in which there is freedom of choice, but no a-prior knowledge of what this choice is about to be. If such a dichotomy were possible, free will would not be such a pervasive problem: it would never clash with knowledge of one's own choice.

This, however, is not the case. Moreover, this could not be the case for rational individuals. First, however habitual a choice is, a rational individual should be able to ask herself whether she indeed wishes to stick to her habit. As soon as the question is posed, the individual will have to admit that she does know her choice, yet that she has a sensation of free will. Second, there

will invariably be intermediate cases, that are not regular enough to require no thought, yet sufficiently regular for the individual to know her own choice.

Rationality requires that we gather information and learn about the environment, our selves and future selves included. Thus, we cannot escape knowledge of certain choices of ours. But rationality also requires that we be able to question these choices from time to time, and this means suspending our knowledge of our own choices. To conclude, free will is an illusion that is inherent to rationality.

4 Another Example[3]

The problem of free will discussed here is similar to a difficulty with the interpretation of strategies in an extensive form game. Such a game is described by a tree, where non-terminal nodes correspond to decisions by individual players. A strategy of a player specifies her choice in *each* of her decision nodes. A solution for the game with n players is an n-tuple of strategies, one for each player. Importantly, such a solution describes what would happen in the game starting from each and every node in the tree, even if this node was not supposed to be reached by the solution under discussion.

Rationality requires that choices be justified by players' reasoning, of the type "I'm planning to do a and obtain the outcome x. If I deviate from a and do b, the outcome would be y, but since I prefer x to y, I should better stick to the choice a". To capture this reasoning, we need to model the players' theories about the way the game will be played, given different choices they can make. The common approach is to take a solution (a strategy for each player) and interpret it as the theory, commonly held by all players, about the play of the game. But then, when a player considers move b instead of

[3]This example may be marginally useful to readers who are acquainted with the debate in the game theoretic literature regarding common knowledge of rationality and the backward induction solution to complete information games. It is probably more confusing than helpful to others.

move a, she knows that she will refute the theory by choosing b. Why should she predict that the theory, which she just refuted, will continue to be a valid description of the play of the game later on? And if her choices causes confusion among other players, how will the player know if such a confusion is in her best interest? How will she evaluated the outcome of her move?

In Gilboa (1998) it is argued that there are three ingredients that make game theoretic predictions inherently problematic:

(i) We wish to describe individual choice. This implies that each player may, single-handedly, refute the theory.

(ii) We wish to focus on rational choice. This implies that the choice can be explained and justified, and the theory can provide answers to questions such as "what will happen if I choose a different move?"

(iii) We seek theories that are consistent with themselves being known by the players.

Relaxing each of these desiderata could obviate the problem. But, taken together, we need a theory of players' reasoning, which can be refuted, and we need to assume that, when a player contemplates refuting the theory, she still believes that she and others will continue to believe in the same theory in making predictions.[4]

The free will problem is described in the context of a single-person decision problem, rather than a game among many players. Still, the two problems are similar in several ways. In both, a rational decision maker knows what the theory predicts she would do. In both, modeling rational choice requires that we consider the decision maker's knowledge (or theory) about the outcomes that would result from her refuting the theory. Thus, in both problems, in order to model the decision maker's reasoning in a logically coherent way, we need to assume that the decision maker somehow suspends her knowledge of

[4]In defense of this problematic assumption, one may argue that a move specified in a decision node only has any meaning if this node is ever reached. Assigning a move to a node, but using it only as long as the node is not reached is tantamount to saying that the players hold theories that are incorrect.

her own choices, while retaining her knowledge about the environment. The ability to know something about oneself, and yet to put this knowledge on hold while making a decision, appears to be necessary for rational decision making.

5 An Evolutionary Story

One can imagine an evolutionary argument for the selection of the sensation of free will, despite its being an illusion. Imagine that there are several species. Species A does not engage in learning at all. Species B learns regularities, and thus knows many of its own choices as well, but is incapable of imagining worlds in which its knowledge is false. Species C can perform learning, buy also hypothetical reasoning, involving counterfactuals.

Clearly, species B and C will do better than A, who fails to predict natural phenomena, the behavior of other animals around it, and so forth. Between species B and C the competition is tougher. Mutations within species B can "experiment" with many decision modes, and the successful ones will survive. Since evolution does the experimentation of species B, no particular individual in it needs to experiment or to engage in counterfactual reasoning. Correspondingly, organisms of species B will follow tradition without experiencing a sensation of free will and with no conceptual difficulties.

However, species C will be more adaptive than B when the environment changes. Assume that organisms of species C are programmed to ask themselves, every so often, "I know that I've been doing a and getting the outcome x for years. Still, maybe it's time to test act b? I know that I do not usually choose b, but why? What would happen if I did?" Only if the organism finds a good answer for not choosing b, will it stick to a. But when such an answer is not compelling, organisms of species C may experiment within their own lifetime, and therefore, facing a changing environment, they will do better than organisms of species B.

The main cost of the counterfactual reasoning endowed to species C will be the need to deal with paradoxes of free will. It appears like a minor price to pay for the ability to make rational decisions.

6 Conclusion

This note does not attempt to resolve the age-old problem of free will. If anything, it attempts to complicate it, by showing that free will is hard to reconcile with the most basic notions of rationality, while also being implied by rationality.

We conclude that rationality is sufficient to explain why we often know our choices, as well as why we have to re-consider them and thereby experience free will. The problem of free will thus does not require any belief in determinism, and it is unlikely to be resolved by scientific research in physics or neurobiology.

References

Gilboa, I. (1998), "Counter-counterfactuals", *Games and Economic Behavior*, **24**: 175-180.

Penrose, R. (1997), *The Large, the Small, and the Human Mind.* Cambridge: Cambridge University Press.

Searle, J. R. (2004), *Freedom and Neurobiology: Reflections on Free Will, Language, and Political Power.* New York: Cambridge University Press.