LUPINE PUBLISHERS

Advances in Robotics & Mechanical Engineering

DOI: 10.32474/ARME.2021.03.000160

ISSN: 2643-6736

Mini Review Article

Assembling Computer with Free Will: An Application of Artificial Intelligence

Yousaf Ali Khan*

School of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

***Corresponding author:** Yousaf Ali Khan, School of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

Received: October 11, 2021 **Published:** October 21, 2021

Abstract

People have a strong perception that they could do otherwise. The perception itself is an experimental fact open to explanation and modelling. There have been centuries of relatively fruitless philosophical debate about abstract concepts of free will. This research work sidesteps the discussion of abstract concepts and focusses on the facts. On the phenomenon of free will and develops the challenge model in particular. This research is an implementation of the challenge model, where an agent has an inclination to assert its independence by responding to an implicit or explicit challenge to do otherwise. A standard utility agent is the basis of the model. A utility function for the agent is derived and applied to a number of free will situations to demonstrate credible performance. To implement a prejudice free test, it is suggested that scenarios are constructed using an alien visitor. We proposed a model of free that will be implemented on an AI system. The manifestation of free will is a consequence of the structure of the utility agent and the value function. It is independent of the physical implementation - it could be biological or a computer. It could be verified using a Turing type test, subject to the specified safeguards. This is a testable scientific explanation of free will.

Keywords: Challenge Model; Free will; Artificial Intelligence; Autonomous Agent; Utility Function; Modelling

Introduction

Humans have a perception that they "could do otherwise". That for a wide variety of actions the individual can make a decision and act in different ways. And that the alternatives are not predetermined. It is a widespread almost universal belief, and it is cross cultural [1]. We call this free will. The perception that we could do otherwise is a fact. There is strong evidence that the perception is based on experience [2] but we lack a scientific model of the decision making. For decades philosophers and scientists have been trying to reconcile free will with current scientific theories [3,4]. Neuroscience has made substantial progress analyzing and understanding the mechanisms associated with our decision making. Scientists are modelling aspects of decision making with causal scientific models. This leaves some tension between the microscopic models and the higher-level concept of free will.

Many, but by no means all, philosophers predict that free will could arise as an emergent phenomenon from underlying processes that are intrinsically causal and predictable [5,6]. A model of emergent free will has not been forthcoming. Philosophers also have an abstract concept of in deterministic decision making. Confusingly this is also called free will. The abstract concept is not evidence based and is untestable [7]. This paper attempts to model the phenomenon. The model stands apart from most philosophical debate for good reason. It results from a scientific, evidence based, approach to understanding human free will. The overwhelming majority of philosophical debate is about abstract concepts with no experimental connection. They make no predictions and cannot be tested by a scientist. One exception in the contemporary free will literature are various investigations aimed at clarifying our folk intuitions about free will and relating the results to differing philosophical views. It's a challenging exercise, see for example [8]. As a scientist intuition needs to be evaluated with care. Intuition that "A" is true may be a fact, but it does not mean that "A" is actually true, it is just a belief statement.

It is worth reflecting that most scientific breakthroughs were counterintuitive at the time. After one hundred years of discussion with no solution in sight, it seems plausible that an eventual understanding of free will will be counter intuitive at first. What we require from a free will model is that it explains the intuition – just as for example special relativity explains the Newtonian viewpoint as a low velocity limit. One might hope that a convincing model of free will would have considerable, even decisive input, to the philosophical debate. We consider a logical structural design running an algorithm. Being an algorithm, it is completely independent of the underlying processes and could be implemented with a variety of materials, including biological cells and neurons, or semiconductor chips. We present a design for a decision-making agent where free will emerges. The design is a logical structural design that runs an algorithm. The design of the algorithm causes free will to emerge.

Own Decision Making

We are not aware of how we make our own decisions. This was demonstrated vividly by Libet's groundbreaking experiments where NMR brain scans purported to show decisions being made before the subject was aware of the choice [9]. Although there is some controversy over what Libet's, and subsequent research proves, what is indisputable is that we don't actually know how we make decisions. The role of subconscious processes influencing, or even determining, our choices are highlighted by the Libet experiments, but it's a factor that was already well known. From unconscious bias in interviewing or law courts, subliminal advertising or impressive stage magicians such as Derren Brown [10] They all show how our reasons for a decision, our conscious decision making, can be completely wrong [11, 12]. We have no awareness of the neuroscientific processes or correlates of decision making. The common perception that our decisions are not predetermined, predates any knowledge or understanding of quantum theory or neuroscience.

Modelling Free Will

It is not unusual to model human decision making. A goal seeking utility agent [13] would be a common approach in economics, game theory, AI etc. The agent has certain goals and takes inputs from the environment. Then some form of rational analysis takes place leading to a decision and an action. The rational analysis works with an explicit or implicit model of the environment and how its actions are likely to produce outcomes. E.g. eating food will satiate hunger. This simple description applies to a wide range of decision-making agents: a basic control system, the battery saver in your smartphone, an insect, an intelligent mammal, a sophisticated AI system, a semiautonomous mars rover, a computer playing chess or poker. The more powerful utility agents will have an element of learning (although that is not always desirable). An extra level of sophistication allows the agent to offer explanations for its action. The architecture is simple, and the functionality is essentially algorithmic. The physical implementation of the utility agent is not specified. It is not important and is not part of the discussion or analysis. This independence is a powerful feature of the model. The common perception of our decision making is not based on, or even informed by, a knowledge of how our brain works. We know if we feel hungry then we want food, and we might work out how to get it. That's an algorithmic not biological or physical description.

We can write an explicit utility function, D, for a Yes/No decision where D is positive for Yes and negative for No. and V_+ and V_- are the utility values for a Yes or No decision respectively. We add a small stochastic element, ϵ, that reflects uncertainties in the decision making. It would be high for someone who behaved erratically.

$$D = (1 + \in)V + \; - \; (1 - \in)V - \qquad (1)$$

The utility values, V, are themselves an accumulation of values for a range of different goals or

utilities that the agent may have. The rational analysis leads to the V values and would normally include balancing different goals or utilities e.g., wanting to eat verses concern about weight.

We can consider a few characteristic distinct scenarios:

- **Tossing Coin**

Choosing heads [or tails] for a game.

V+ = V- and both are small values

D is almost zero. The actual decision is affected by the ϵ factor giving a 50/50 outcomes

- **Committing Murder**

There may be some advantages to you killing X, but the adverse consequences are thankfully overwhelming

V- >> V+ and V- has a large value

D is always negative. There is such a big difference between V+ and V- that ϵ cannot change the outcome and the decision is N. Murder is rare.

- **Writing with your right –hand**

Choosing which hand to use to write your signature:

V+ >> V- and but quite small values

D is always positive. Although V+ and V- are small there is such a big relative difference between V+ and V- that ϵ cannot change the outcome and the decision is Y. The agent uses its right hand.

- **Destructive Acts**

Taking some action that is harmful to the agent, with no benefit.

V+ =0. V- could range from small (a pin prick) to as large for an action that results in death.

This action would not be performed, there is no value to it. The utility agent immediately offers a model of a decision-making system that is powerful and successful at reproducing many aspects of human decision making. But what is missing? There is nothing identifiable as free will. We need something else. Some obvious answers are wrong! Predictability and unpredictability are already included. A chess playing computer will predictably make moves consistent with the rules of chess. Some moves may be so obvious to satisfy the goal of winning, that a good chess player could predict them. While other choices might be surprising – that is how it wins. Even more so for an AI poker playing agent! There could also be an element of randomness – agents controlling network traffic, may use a random delay to avoid repeated traffic conflicts.

Randomness could be implemented simply by looking up a number in a preexisting table, or the microsecond reading of a clock, but it could equally well be the output from a radioactive decay process. Our best description of the latter is quantum theory – which is in deterministic. So, the agents can have a mix of predictability and unpredictability, they can make different decisions in similar circumstances.

We need to consider if the agent "could do otherwise" or to be more precise does the agent have a perception that it could do otherwise. And as objective observers, do we see any evidence that it could to otherwise. If not, then is a modification possible to reproduce the phenomenon of free will? We could consider finely balanced decisions, ones that appeared to be 50:50 choices and ask: could the agent have chosen otherwise. But that is not helpful. A small change of circumstances, or a small stochastic element of the decision making could affect the outcomes. In similar situations a different choice may be made, but that is an unremarkable feature of even the simplest of decision-making units. A deliberative analytical process might correctly describe two evenly weighted options, both equally advantageous in achieving the goals. Again, that is unremarkable.

To look for free will we need a highly discerning test [7]. Consider a decision that is quite predictable, where the converse does not meet any of the agent's goals as described above and might even be harmful That would be example (D) above. Could an animal avoid food if it was very hungry, or a robot put itself in danger, near a flame for example, or a game player making a losing move (not a bluff). A highly discerning test is to challenge a highly predictable decision to do otherwise. The test outcome is a decision, and action, that would not, as far as be known, meet any of the agent's goals. If we apply a highly discerning to test to any of the agents described above, they will fail it. Given the challenge as an input and the goals they have to satisfy; they will not respond to the challenge - why should they?

A human could respond to the challenge, a human could do otherwise, but a utility agent could only respond if the response was judged to advance one of their goals. To model human free will we need to add a goal to the agent that is satisfied when it responds to a "Could you do otherwise?" challenge. Of course, this will be one of a number of competing goals. We call the new goal independence. The agent asserts its independence by responding to a challenge to do otherwise. A prediction of behavior would be one such challenge ... "you are going to ..." We can add an extra term to the utility function I, which is positive if yes satisfies independence and negative if it undermines independence. The magnitude of I will rise in response to a challenge.

$$Dfree=\left(1+\in\right)V+-\left(1-\in\right)V-+I \qquad (2)$$

Applying the new utility function explained in Eqn.2 Dfree to the examples above has varying degrees of effect

- **Tossing Coin**

Choosing Tails [or Heads] for a game.

V+ = V- and both are small values. I dominate, but because it was a 1/2:1/2 choices without the challenge, the challenge effect will not be demonstrated by one decision but should show up in the statistics of repeated decisions with and without challenges.

- **Committing murder**

V- >> V+ and V- has a large value. V+ dominates I Behavior is most unlikely to change. You cannot easily challenge someone to commit murder.

- **Writing with your right –hand**

Choosing which hand to use to write your signature:

V+ >> V- and but quite small values. Because V+ and V- are small I will dominate. This is a highly discriminating test evident in a single decision. The choice will switch from Yes to No. A previously highly unlikely outcome will become the most likely.

- **Harmful Acts**

Taking some action that is harmful to the agent, with no benefit. V+ =0. V- could range from small (a pin prick) too large for an action that results in death. Depending on the magnitude of V-, I could overcome it and change the outcome. This is another highly discerning test. A very unlikely outcome, of no apparent value, become more likely and in some cases most likely. Where V- still dominates I there will be no change – a decision that entails serious harm or death is most unlikely. An interesting feature of the challenge stimulus is that it need not be external. In principle the agent could generate a challenge to itself.

Possibly will do Otherwise

We have now designed an agent that could do otherwise. For any decision, the agent could be challenged, and that challenge would affect the outcomes. On balance, the probabilities would change. For highly discerning tests the change of outcomes would be clear and otherwise inexplicable. An agent that could generate its own challenges, would also see a change of outcomes. It would know that for any decision, it could raise a challenge and potentially change the decision. It would see it responding to challenges from outside. Over time it could build a historical record of decisions and alternatives choices, where a challenge switched from one decision to another. It explains the outcomes reported by experimental philosophers [2,8] on perceptions of our free will. A goal seeking agent has some element of modelling or predicting the future simply to evaluate the effect of a decision against a goal it seeks. An independent agent would know that for any upcoming decision it could do otherwise. A crucial feature of the agent's perception and analysis is that it is unaware of the detailed decision-making process, it is unable to predict when a challenge or a challenge to a challenge might arise, but it does know that if it initiates a challenge there is a consequential change to the decision making. Consequently, with the knowledge the agent has, the future is not predetermined.

Conclusions and Suggestions

The decision-making properties emerge from the algorithm not from the physical or biological implementation. It is well known that even small simple algorithms can have unpredictable outcomes, so this is perhaps not too surprising. In this case the property that we call free will emerges through the choice of programmed goals [independence] and stimuli that the agent is sensitive to [challenges]. With the knowledge that the agent and observer have, the decisions are not predetermined. Arguably no amount of knowledge could change the appearance of free will because the knowledge itself is a new input - a type of challenge. So internal knowledge adds a further recursive aspect to the algorithm. However, an external observer, may well understand the algorithm well enough to make secret predictions that the agent is unaware of. The observer may realize that the decisions are predetermined or can be controlled. Such insight is a well know feature of stage magicians, parents managing their children and indeed many other aspects of interpersonal relations.

This is a model of human free will - the phenomenon that we know we could do otherwise. It becomes more powerful when the agent has a capability of self-awareness and handling abstract concepts. With self-awareness, the agent knows its history of decision making and challenges; it knows it could do otherwise. Note that this self-awareness does not and cannot extend to the inner workings of its analytical engine. Given an ability to handle abstract concepts the agent would apply the terms free will and indeterminism to its decision making. The single extra goal and one-dimensional challenge stimuli is over simplistic for human decision making. We could, for example, consider a challenge as an active curiosity. “What if I did something different?” leading to different decisions and actions. In reality human goals, stimuli and responses are far more complicated. However, adding this single extra goal of independence directly, immediately and unambiguously creates an agent that could do otherwise. Which is not a bad starting point for a resolution of the free will problem.

This model makes predictions and is testable. An agent that does not respond to a challenge will not be perceived to have free will. An agent that cannot respond to a challenge will not have the perception of free will. You can imagine these scenarios yourself. However objective testing is difficult because free will is associated uniquely with humans who have many other particular characteristics, including consciousness (in many forms, defined or otherwise!). Free will as a phenomenon can be difficult to extract and is vulnerable to many prejudices. One route to objective tests is through fiction and even better cartoons, where form and behavior can be presented independently. Take Spock from Star Trek, portrayed as human in appearance, but with entirely logical decision making, he cannot do otherwise if it is illogical. At the other extreme, is Bender in Futurama. Bender is drawn as the simplest of cartoon, tin can, robots, but behaves with independent decision-making behavior as complete as the other humanoid characters. Bender has deviant behavior, but it responds to challenges, and generates its own challenges. Do viewers regard it as having free will? Episodes could be written specifically to provoke and analyze attributions of free will.

One fascinating approach to testability of this free model, and any others, is a Turing type test. If we consider an alien visitor in a suit. We cannot tell if it is an automaton, or a free agent like us. It is an alien, so we can’t rely on intuition. How do we decide if it has free will? The answer is to do a challenge test [7]. Turning the test round. We can propose alternatives descriptions of the alien behavior and which if any are perceived as those of a free agent. These are the techniques used by experimental philosophers. The model also explains false attributions of free will. Some peoples attribute free agency to natural phenomena like weather or volcanism. It is a known human characteristic to find patterns even when they do not exist. People can imagine a pattern of response to challenges and consequently assign agency and make offerings to appeal to the free will of the spirit.

Declaration

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

This research is financially supported by National Natural Science Foundation of China under Grant.No.71961127.

Availability of data and material

Research code and data used in this research will be available on request upon acceptance of this research.

References

1. Sarkissian H (2010) Is Belief in free will a Cultural Universal? Mind and Language 25(3): 346-358.
2. Deery O (2015) Why people believe in indeterminist free will. Philosophical Studies 172: 2033-2054.
3. Kane R ed (2002) The Oxford Handbook of free will. Oxford University Press, Oxford.
4. Searle J R (2007) freedom and Neurobiology: Reflections on free will, Language, and Political Power. Columbia University Press.
5. Dennett, D (1984) Elbow Room: the varieties of free will worth wanting. Clarendon Press,
6. Smilansky, S (2000) free will and Illusion. Clarendon Press, Oxford.
7. Hadley, M (2018) A deterministic model of the free will phenomenon, Journal of Consciousness Exploration and Research 9(1).
8. Nadelhoffer, T and Yin, S and Graves, (2020) Folk Intuitions and the Conditional Ability to Do Otherwise. Philosophical Psychology 33(7).
9. Libet B (1985) Unconscious cerebral initiative and the role of conscious will, Behavioral and Brain Sciences. 8: 529.

10. Brown D (2006) Tricks of the mind. Channel 4 Books. Oxford.

11. Stanovich K (1986) How to think straight about psychology. Scott Foresman.

12. Nisbett R and Ross L (1980) Human Inference: strategies and shortcomings of social judgement. Prentice Hall Englewood Cliffs N J.

13. Russell SJ and Norvig P (2013) Artificial intelligence: A modern approach. Prentice Hall Englewood Cliffs N J.

DOI: 10.32474/ARME.2021.03.000160