Mind-forg'd Manacles and Somewhat Free Will

Steven Brock
stbrock@stbrock.net
April 24, 2020

In every cry of every Man,
In every Infants cry of fear,
In every voice: in every ban,
The mind-forg'd manacles I hear

—William Blake, *London*

Abstract

Free will is reconceived in terms of degree of unpredictability rather than absolute opposition to determinism. So conceived, it retains its traditional expected character: unpredictable choices which in retrospect reflect individual traits and justify moral responsibility. This "somewhat free will" is explained in terms of counterpredictive strategies, restrictions on computability by embedded agents, and self-referential interaction between the unconscious and conscious mind. Libet's finger moving experiment, widely regarded as the strongest such evidence against conscious will, is reconsidered and appears consistent with this view of free will.

Somewhat Free Will

Conceived from a logical perspective as an entity that must either exist or not exist, the humanistic principle of free will seems irreconcilable with the grim doctrine of determinism. Yet with the astonishing advances in the physical sciences since the Renaissance, determinism has gained a scientific aura, while free will, if still cherished, has come to seem unscientific. Reconceived in terms of degree of predictability rather than deterministic computability, free will becomes compatible with determinism while retaining its traditional characteristics to the degree it is present. Choices are not fully predictable in advance, even in theory, but remain consistent with a scientific worldview. Choices in retrospect are not random, but arise to a degree out of individual characteristics. Choices offer traditional justification for moral responsibility and social principles.

The analysis begins with a strategy for complete unpredictability by a hypothetical predictive Turing machine making revealed predictions, then extends that strategy to supposedly unrevealed predictions and in degrees to a full range of commonplace circumstances commonly believed to involve free choices. Free will so conceived can arise as the mind seeks to model and interact with its environment, including itself and perhaps other free agents. It can also arise from the self-referential interaction between an individual's unconscious and conscious mental processes.

Turing Predictions—Revealed and Unrevealed

To investigate the interaction of prediction and somewhat free will, we will enlist Alan to predict whether or not Emily will lift her finger using what we will call a predictive Turing machine, which is assumed sufficient to predict with total accuracy the relevant environment and the lifting of the finger. He checks the display and announces his predictions, one by one. Emily does the opposite every time.

Why isn't this already enough for a degree of free will? Emily makes choices that can be seen as an expression of her individual character. They are hardly random, following an obvious pattern. Yet her predictably unpredictable choices defeat every prediction of the predictive Turing machine.

The usual argument to the contrary is that concealed predictions remain reliable and preserve a deterministic worldview.[1] Rerun the same scenario with Alan and his predictive Turing machine behind a one way mirror making predictions while Emily, unaware, happens to lift her finger occasionally. Alan steps out from hiding, shows her the videotape of accurate predictions, and says "Thus, you are an automaton." She slaps him. He takes from his pocket the prediction of the slap.

This seems at first like a plausible philosophical argument against free will in a deterministic world, but on reflection it actually depends on physical force to succeed. If Emily wanted to look behind the mirror or go to the predictive Turing machine store and get one for herself, Alan would need to manacle her to a tree to prevent it. If Emily had her own predictive Turing machine she could self-predict and then still act to the contrary, to some degree freely in the sense of unpredictability. Denying her access to the same self-predictive methods as Alan undoes the argument that her actions are predictable and not free. Alan would in effect be cutting off her finger and then predicting that she wouldn't lift it.

There is the further problem that the Turing approach to these circumstances consistently yields either wrong predictions or no predictions at all.[2] That is not how things actually happen. Attempts to avoid these results simply switch to a different but equally unsuccessful cycle of avoidance, counter, counter-counter, and so on, with Alan either making a final prediction which Emily finally and decisively counters or falling into infinite loops and predicting nothing at all. If subroutines to recognize and break out of the loops are attempted, they will likewise be incorporated in the predictive Turing machine's assumed accurate predictions—and then countered. If Alan's machine were ever to halt and return a prediction, it would be wrong. So it will keep on looking for a way to make a prediction which it can't anticipate being countered, which never happens: infinite regress with no prediction forthcoming.

Recasting the situation in terms of game theory reaches equivalent results.[3] Even if Alan has a vaporware oracle Turing machine, Emily can counter with her own equally capable oracle machine. Again, they come to an impasse. Or perhaps she may let him get it right every now and then.

The striking thing about these outcomes is not so much the apparent limitations on computability and prediction as the fact that the results of the algorithmic modeling of human behavior are not just flawed, but either wrong in every case or infinite loops—neither of which are the sorts of

---

[1] The problem with failed prediction in the face of counterpredictive strategies has received some attention, but relatively little recently until [Ismael 2019].

[2] Lloyd's Turing test of free will is actually only a computational rationale for subjective human belief in free will—a questionable solution for a non-existent problem. [Lloyd 2012]

[3] The somewhat adversarial moves and countermoves between Alan and Emily, each as independent strategizers, suggest that game theory might provide a resolution where computational theory could not. [cite Suppes] Prediction and counter-prediction is a poor fit into the 2-person zero sum game mold, but could be recast in various ways. Alan and the predictive Turing machine would always move first. With moves in the open, Emily could simply counter and win each time. With moves hidden, Emily would by assumption be predicted and lose without her own predictive Turing machine but win with it. The results are essentially unchanged, with Emily only losing in the suspect alternative where she is denied her own predictive Turing machine and the moves are hidden.

outcomes encountered in everyday human experience. We're rarely systematically wrong continually and never fall into infinite loops. This suggests that the difficulty is not a matter of algorithmic insufficiency, or the difference between revealed and unrevealed predictions, and perhaps goes beyond issues of computation and prediction.

The irony is that when Emily counters, or doesn't counter, Alan's prediction, her decision becomes predictable after the fact. To the extent that she herself predicted, hypothesized, or simply guessed what she would do, she could have countered that self-prediction as she could Alan's. But once she acts, the chain of explanation is fixed.

Ismael provides a history of the work on counterprediction and determinism as well as a careful consideration of related issues of predictions and counterpredictive devices and argues that the ability to counter revealed predictions is in itself sufficient to establish that it is impossible to predict from within the universe its exact unfolding over time from its initial state, though scientific determinism in some sense is preserved from an outside perspective. [Ismael 2019]. Ismael's outside perspective is developed in terms of light cone limitations on causal effects in Minkowski spacetime. It may be helpful to imagine this in terms of Einstein's vision of the universe from outside as a fixed four-dimensional "block" of spacetime, though Ismael does not do so. [Ismael 2019] Ismael does not however draw conclusions on free will, having previously developed a different viewpoint. [Ismael 2016]

Here we will argue that limits on predictability as described can arise from the uncomputability of any environment by an agent embedded within it and also from the self-referential interaction between the conscious and the unconscious mind. From this perspective, it appears that unpredictability potentially arises, at least to some slight extent, in almost all human actions, excepting those that are physically impossible.

Analogs of Prediction

From the perspective of somewhat free will, there are many situations which from a linguistic standpoint are quite similar but yet are not explicitly predictive and do not seem so potentially threatening. The imperative "lift your finger" could be countered just as systematically as was the prediction, as any two-year old well knows. The same scope for some mixture of causal pressures with some residual freedom to comply or not seems likely in the abstract and common in experience. Another step takes us to "lifting a finger is required by law," which is close already to establishing at least a degree of free will to support moral responsibility and deterrence, primary focuses for interest in free will.

Language of belief, desire, intention, evaluation, likelihood, and on and on, all have functions comparable to the language of prediction. Simple factual statements about anything can be interpreted as predictions of experience likely to be had in some context, as could statements about the past, as well as the present and future. Beyond language, physical processes and objects can be interpreted by humans as predictions of their changing or persistent natures. Turing was well aware of these possibilities and once told Popper that a blank sheet of paper could be interpreted as a prediction that it would persist in its state as long as it wasn't interfered with. [Popper 1982]

The situations underlying these common language usages vary in the degree to which they may put causal pressures on everyday choices, but all would seem to leave room for some degree of unpredictability in response. The psychological power of these pressures, and even whether they

would make certain choices more or less likely, seems quite difficult to predict and could itself become unpredictable if the person making the choice takes it into account. While we may not have the capability to estimate such unpredictability with any precision, the existence of factors which could contribute to unpredictability may be commonplace and uncontroversial. Choices which appear somewhat unpredictable will involve some potential freedom of choice and thus some responsibility for the choice.

Sources of Unpredictability

Counterprediction as demonstrated by Emily and Alan is one source of unpredictability. There are two common processes which seem particularly liable to generate unpredictability in similar fashion: embedded agency and self-reference, particularly self-referential thoughts.

Embedded agency is one of several quasi-computational processes which simply do not always yield a single, exact result. Thus, they are generally not functions and will not be Turing computable. Interactive computation, for example, aspires to model complex systems such as the internet allowing parallel computations and continual input and output. [Goldin 2006] However, the task still appears daunting, and the internet continues to run on a backbone of Turing computation, with programs constructed to keep out of one another's way most of the time and to recover most of the time if they do crash.

Artificial intelligence researchers are seeking a theory of computation for agents interacting in a shared environment where all need to predict each other's actions, thus far with limited success. Current approaches are in effect heuristic, as there is not yet general agreement on how to formalize the research for an underlying theory—which remains for now "a mystery." [Demski 2019; Soares 2015]. However, the less familiar and perhaps more significant source of uncertainly in prediction may be self-reference in the human mind arising out of the interaction between its conscious and unconscious processes.

Self-reference

If internal unconscious psychological and other factors push a suggested action into consciousness and it is acted on without consideration, as an impulse perhaps, there is no opportunity for self-reference. But thinking about a thought creates a degree of free will arising out of the self-reference by creating the opportunity to accept or counter it, as in the Alan and Emily scenarios, or respond in other more complicated ways. The choice made after consideration of the unconscious mind's proposal is less predictable because it has available additional information: the recommendation from the unconscious. Similarly, an action taken after knowledge of a Turing machine prediction of your action is based on different data than a decision without such a prediction. The process of thinking further based on additional information simply creates the potential for a choice different from that which might have been made before. Internal thought generation and processing may often proceed by relatively routine and consistent processes or heuristics and generate little reconsideration or likelihood of changed action. But external information bearing on our actions may generate more substantive reconsideration.

Many thoughts that are not expressly predictive create the same opportunities for confirmations, counters, or any variety of other responses. Imperatives, laws, suggestions, statements, even reactions to an object or experience—all may change the information available and possibly, if only slightly, weigh on a decision being made.

The sequence of actions—first, external prediction, second, internal review including the new prediction, and third, decision taking external prediction into account prior to decision—makes vivid why the external prediction can be confounded, especially in situations where differences between options are minimal. A choice might be affected, the probabilities changed a little or a lot. A deterministic (if that is what it is) process, is available to the one making the decision, but only after the choice is made.

An attempt to backport the internal reaction to the external predictor would simply create an infinite looping of counteractions. In modeling the potential series of predictions and counters to predictions that could result, the external prediction could be changed in any manner and the internal recipient could choose to confound it again. And if the external predictor included a subroutine to monitor for and exit its looping and make another prediction, the recipient could simply confound it as before.

There is no way for the external predictor to force a correct prediction. Yet the recipient's actions may be determined and predictable in themselves from another perspective. And the recipient can fashion its responses so that, one by one, they are unpredictable. Our lack of understanding of our mental processes and the fact that only small amounts of time can be devoted to most decisions combine to avoid the theoretical back-and-forth of counters in everyday affairs, but a degree of free will is inherent in the possibility and is sometimes asserted.

Attempting to understand our conscious mind with our conscious mind is inevitably hazardous, and likely impossible at least in part. It becomes harder still when the conscious mind treats its input from the unconscious as if it were somehow its own creation. It is as if the unconscious were behind a one-way mirror, observing conscious activity, working on conscious concerns in its own way, proposed answers on the mirror for conscious consideration—without showing its work.

The interaction is in a broad sense like that between Alan and Emily, with the unconscious processing experience, calculating responses of all sorts based on heuristics, biases, evolutionary predispositions—a variety which we are only beginning to investigate. The conscious mind can then respond and proceed in ways that are unpredictable in advance but explainable in retrospect—with a will somewhat free, yet still arising out of individual character.

The conscious/unconscious bifurcation of the mind provides the structural underpinnings for unpredictability. We can view the mind as two agents embedded in a shared environment, each having to model itself as well as the other to understand and control its world—a classic example of unpredictability which seems to be becoming more and more crucial to the development of what we call artificial intelligence. Or we can view the mind as inherently and continually self-referential—Emily's consciousness continually receiving Alan's computed predictions and freely making decisions predictable only in retrospect.

Libet's Finger

Against these computational and philosophical considerations, there is Libet's much discussed experimental work often taken to minimize or, according to Daniel Wegner, eliminate any ability for conscious free will to make genuine choices. Libet's elegant experimental design monitored brain and muscle electrical activity leading up to "spontaneous" lifting of a finger. The results, as described by Wegner:

> suggest that the brain starts doing something first (we don't know just what that is). Then the person becomes conscious of wanting to do the action. This would be where the conscious will kicks in. . . . Then, and still a bit prior to the movement, the person reports becoming aware of the finger actually moving. Finally, the finger moves.

The implication is that the moving of the finger originated in unconscious brain activity prior to any conscious choice to do so. The unconscious "something" that the brain was doing was generating a readiness-potential in the nervous system, necessary preparation for both voluntary and involuntary movement, which in this particular experiment was closely coordinated with the following finger movement. The readiness potential was considered simply "an indicator of cerebral activity" in the unconscious, possibly initiated earlier in other regions of the unconscious. Libet's conclusion was unequivocal:

> It is clear that neuronal processes that precede a self-initiated voluntary action, as reflected in the readiness-potential, generally begin substantially *before* the reported appearance of conscious intention to perform that specific act.

The specifics of Libet's results have been subjected to extensive scrutiny and efforts at replication and extension, with mixed opinions on the results.[4] In *The Illusion of Conscious Will*, Wegner interpreted Libet's results to mean that conscious will was a psychologically useful illusion, but played no role in actual causation of behavior. [Wegner 2018] Libet himself thought the conscious mind had a narrow window to veto the unconsciously initiated action. Here, the experimental results will be assumed correct, but will be viewed from a different perspective and found plausibly initiated by prior conscious processes after all.

Consider how you take your keys from a hook by the door as you walk out to go to work. You make a point of putting them back on the hook when you return, so it becomes reliable, habitual, for your hand to take the keys, leaving you free to be thinking about work already as you walk out the door. Until one day you're interrupted because your hand is fumbling at the hook. The keys aren't there, and conscious attention is required to find them. This unconscious capacity to remember to do something in some future context is called prospective memory. The most studied forms of prospective memory involve triggering of action by an event or passage of an interval of time and typically involve explicit one-time intentions. [McDaniel 2007; Bradimonte 1996] However, habitual prospective memory tasks, such as taking the keys going out the door, are different, and "the intentionality of habitual tasks may be only implicit in the overall task." Pilot training, for example, establishes habitual prospective memories to trigger automatic performance of particular sequences of checks in particular situations. [Dismukes 2012].

Libet's experimental design falls comfortably within the habitual prospective memory paradigm, with training of participants comparable to pilot training: extensive conscious preparation and training of the finger motion and a clear conscious intention in advance for the finger to move during a specific short interval of each run of the experiment. The subjects were six college students, who had one or two half-day training sessions, and six to eight half-days of experimental runs with additional training runs at the outset of each. Each participant performed hundreds of

[4] [Saigle 2018; Pacherie 2011]. Pacherie argues generally that higher level conscious intentional processes akin to prospective memory could have initiated the moves of the finger but does not consider the direct relevance of habitual prospective memory to the experimental design and results.

experimental runs over the course of the experiment. The series of events being measure took place within 1–2 seconds, so for timing accuracy the subjects were trained until they could flex their hand consistently in less than 20ms. (While the movement is generally referred to as moving or lifting a finger, it appears to have been more of a flexing of the hand.)

The subjects were generally instructed:

> "to let the urge to act appear on its own at any time without any preplanning or concentration on when to act" that is to try to be "spontaneous" in deciding when to perform each act; this instruction was designed to elicit voluntary acts that were freely capricious in origin.

How much time was available to be "spontaneous" and "let the urge to act appear on its own at any time"? Literally, an eye blink. In the interest of precise timing, each run had to be completed without blinking.

The supposed roles of the conscious and unconscious mind during this brief interval seem confused and contradictory. How are we to understand that a verbal instruction of spontaneity to the conscious mind was effective when the conclusion of the experiment was that the conscious mind did not initiate the lifting of the finger at all? If the movement was initiated unconsciously, how was that accomplished? The readiness-potential is part of the muscle flexing neural structure, not part of any identified decision-making area in the unconscious.

The rationale for the Libet interpretation seems to be that no conscious source was found during the brief interval of the experimental runs, so whatever the trigger of movement was, it had to be unconscious, like the readiness-potential. That rationale overlooks the possibility of conscious training of implicit intention to move during the prescribed interval in the experimental design—but formed *prior* to that interval and executed more or less automatically by the unconscious when the time came.

That is what analysis of the Libet experiment in terms of habitual prospective memory provides. Without experimental data locating the origin of the "spontaneous" finger movement in the unconscious or excluding the operation of habitual prospective memory, the Libet experiment can remain a classic demonstration of internal brain function, but not of the absence of free will.

The widespread inability of either side in the debate over the implications of Libet's experiment to look beyond the time frame of the experimental run is a striking example of the mind-forg'd manacles that limit the conscious mind's awareness of its inextricable entanglement with the unconscious.

Gutei's Finger

The *Mumonkan*—Gateless Gate—tells the story of Gutei, the master of one-finger Zen, which, in an unreliable amalgam of half-remembered bits from anonymous translations, perhaps goes something like this:

> Old Gutei lived in a small village, perhaps much beloved, perhaps not, and taught his Zen to whoever would listen by raising his finger whenever he was asked a question

> and never saying a word. One day a boy who studied with Gutei was asked what Gutei taught and the boy raised his finger. Gutei immediately summoned him and cut off his finger. As the boy ran away crying, Gutei called for him to stop. The boy looked back. Gutei raised his finger. The boy was instantly enlightened.

The answer is not that Gutei was pointing upward at the daytime sky—a one way mirror through which the universe observes us unseen. Nor is it that the starry nighttime sky is the reality behind the mirror, since all you see there is a billion years out of date. Nor is it that you are imprisoned here on Earth until you calculate where to exit.

Conclusion

The new scientific program of neurophenomenology seeks to assimilate our knowledge of neuroscience with our experience of consciousness through study of the embodied mind. If eventually successful, it will become something akin to a mechanics of consciousness and behavior. We gain a glimpse of its boundaries when Emily demonstrates she can almost effortlessly frustrate the predictions of any possible Turing machine—from inside. Yet from outside, her thoughts and actions would seem as predictable and determined as her past would seem to be to us if we had the tools to monitor it or to a predictive Turing machine predicting her past up to the moment before now. The multitude of analogs in our experience where self-reference seems to reset some progressions to zero for a fresh start foreshadows a wide scope for this unpredictability.

While the science of neurophenomenology may seem for now "mysterious" at best and "soft" at worst, the fault is not inherent in nature and simply reflects the inadequacy of our current science. Understanding the causal forces at work in the mind and the operation of an unpredictability within it that we may as well call a somewhat free will can help us predict and (hopefully) constructively direct our actions as individuals and societies.

Our actions seem to us as free subjectively as they seem inevitable scientifically. There appears to be a little truth to both. Sorting out and gaining some measure of understanding and rational influence over our human choices, our somewhat free will, is at least as crucial to the future of humanity as relativity and quantum mechanics.

We all know that there remain mysteries at the core of relativity and quantum mechanics. We have learned to forget them during the week as we measure, compute, and create with whatever seems to work—then worry about the mysteries on the weekend. We don't yet know enough to harmonize these somewhat incommensurable paradigms. But we should know enough by now to live and let live with the contradictions while they help us create better science.

As Emily said:

> God of the Manacle
> As of the Free -
> Take not my Liberty
> Away from Me –
>
> —Emily Dickinson, *Let Us play Yesterday*

BIBLIOGRAPHY

1. Bradimonte M, Einstein GO, McDaniel MA, Prospective Memory, Theory and Applications. Erlbaum Assoc. 1996.
2. Demski A, Garrabrant S., Embedded Agency. Berkeley, CA: Machine Intelligence Research Institute 2019. arXiv:1902.09469v1
3. Dismukes RK, Prospective Memory in Workplace and Everyday Situations (2012). Current Directions in Psychological Science, 21(4), 215-220.
4. Dickinson, E, The Poems of Emily Dickinson. Edited by R.W. Franklin. Belknap Press, 1998.
5. Goldin D, Smolka SA, Wegner P, Interactive Computation. Springer 2006.
6. Ismael J, Determinism, Counterpredictive Devices, and the Impossibility of Laplacean Intelligences (2019). The Monist, 102, 478-498.
7. Ismael J, How Physics Makes Us Free. Oxford 2016.
8. Libet B, Gleason CA, Wright EW, Pearl DK (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.
9. Lloyd S, A Turing test for free will (2012). Philosophical Transactions of the Royal Society A 370, 3597-3610.
10. McDaniel MA, Einstein GO, Prospective Memory. Sage 2007.
11. Pacherie E, Haggard P, What are Intentions? (2011). In Sinnott-Armstrong W, Nadel L, eds., Conscious Will and Responsibility, 70-84. Oxford 2011.
12. Popper KR, The Open Universe. Routledge 1982.
13. Saigle V, Dubljevic V, Racine E, The Impact of a Landmark Neuroscience Study on Free Will: A Quantitative Analysis of Articles Uing Libet and Colleagues' Methods (2018). AJOB Neuroscience, 9(1), 29-41.
14. Soares, N, Formalizing Two Problems of Realistic World-Models. Technical report 2015-3. Berkeley, CA: Machine Intelligence Research Institute. https://intelligence.org/files/RealisticWorldModels.pdf
15. Wegner, DM, The Illusion of Conscious Will. MIT Press 2018.