Protean Free Will

Liam F. Clegg*

March 28, 2012

Abstract

This paper argues that free will is a purely theological issue, and offers an error theory for the free will debates in analytic philosophy in terms of evolutionary naturalism. I introduce 'protean free will' (PFW) as the ability to play mixed strategies effectively in noncooperative interactions. Thence, I argue that traditional worries about divine foreknowledge, Frankfurt controllers, moral responsibility, and determinism are side effects of selective pressures for unpredictability in our evolutionary past. Finally, I interpret the Libet experiments as showing an adaptive response to such pressures. I conclude that PFW does most things most philosophers want free will to do, conditional on the nonexistence of God.

1 Introduction

Free will is a perfectly meaningful metaphysical property in theology. It can be invoked, for instance, to address the apparent quandary posed by a God who is omniscient, omnipotent, and omnibenevolent allowing people to

*Division of Humanities and Social Sciences, California Institute of Technology. Special thanks to Daniel Dennett and the participants in his Free Will seminar at Tufts University. Thanks also to Kenyon Bradt, Tristan Chambers, Simon Reid, and Margaret Smith for helpful feedback on earlier drafts.

suffer and condemning some souls to eternal damnation (e.g. Lewis, 1940). There are a number of ways free will could exist in this context, and so it is potentially worthwhile and productive for theological philosophers to debate the existence and meaning of free will. After the progress in theology in the last half-century, though, free will is no longer a particularly compelling issue in most theological circles.

Yet free will as an issue of debate has enjoyed a resurrection in recent decades among philosophers working in the analytic tradition. Unsurprisingly, these philosophers have made little meaningful progress. For like young-earth creationists studying paleontology, they have thrown out the Baby and are perplexed by the bathwater that remains.

If free will is properly a theological concept, then it may be meaningfully approached in either of two ways: theologically, according to the methods of that discipline; or naturalistically, as Barrett (2000) and Dennett (2006) approach the more basic theological concepts. I offer a naturalistic account of free will and the free will debate.

The paper proceeds as follows. Section 2 defines Protean Free Will (PFW). Section 3 compares the relationship between free will and divine foreknowledge to that between PFW and Frankfurt controllers. Section 4 shows how the sort of alternative possibilities entailed by PFW support moral responsibility. Section 5 discusses the implications of determinism and indeterminism for PFW. Section 6 derives the physiological requirements for PFW, and Section 7 offers an interpretation of the Libet experiments which

shows how these requirements might be met by the human brain and mind. Section 8 compares PFW with other compatibilist notions of free will, and some incompatibilist ones too. Section 9 concludes.

2 Protean Free Will

Consider an agent who faces an environment which includes sophisticated other agents with interests contrary to hers. Call the agent Mary, and call the other agents predators. One good way for Mary to avoid exploitation by predators, exploitation which may include death, is to engage in *protean behavior* (Chance and Russell, 1959). That is, she may behave somewhat erratically so as to be unpredictable. As documented by Miller (1997), the protean strategy offers many clear advantages over the alternatives of concealment of intentions and active deception. While most notions of 'reason' prescribe a single optimal action in any situation,[1] Mary's behavior must sometimes be locally sub-optimal for the sake of unpredictability.

[1]Descartes (2008/1641, Meditation IV, p. 58): "if I always clearly saw what is true and good, I would never need to deliberate about a judgment to be made or a course of action to be chosen; and in that case, although I would be fully free, I could never be indifferent."
Locke (1985/1689, II/XXI: Sec. 48): "A perfect Indifferency in the Mind, not determinable by its last judgment of the Good or Evil, that is thought to attend its Choice, would be so far from being an advantage and excellency of any intellectual Nature, that it would be as great an imperfection, as the want of Indifferency to act, or not to act, ill determined by the *Will*, would be an imperfection on the other side."
Kant (1956/1788, p. 31): "The practical rule, which is thus here a law, absolutely and directly determines the will objectively, for pure reason, practical in itself, is here directly legislative."
Samuelson (1947, p. 111): "[The demand functions] must be single-valued; i.e., to each set of prices and income there corresponds a unique set of goods."

Such local sub-optimality means that at some times t, there are multiple courses of action A_t available which are tied for the strategic optimum. Call the set of such courses of action Mary's strategy S_t. For the purpose of this discussion, no generality is lost in assuming that the optimal probability distribution over S_t is uniform, so that each A_t is equiprobable. Furthermore, while S_t is defined from a bird's-eye view, we may assume that Mary has evolved some reasonably good mechanism for approximating it in the real world. The elements of S_t depend on Mary's preferences, of course, but the strategic optimality of each course of action also includes the risk of exploitation by predators. This, in turn, depends on Mary's past actions and predators' resulting guesses about her next action. If Mary and her predators both use optimal mixed strategies (von Neumann and Morgenstern, 1944), each A_t should be equally optimal for Mary, and there should be no advantage to a representative predator of predicting that Mary will perform any given $A_t \in S_t$ rather than any other $A'_t \in S_t$. However, if a predator knew *or reasonably suspected* at time r, $r < t$, that Mary were going to perform A^*_t at time t, the predator could exploit this knowledge. A^*_t would therefore no longer be optimal for Mary, and would therefore not be an element of S_t when time t arrived. Call the ability to select an A^*_t from S_t and perform it such that A^*_t is still in S_t at time t 'protean free will' (PFW).

3 Divine Foreknowledge

God's omniscience has long been seen as a potential impediment to free will. Augustine caricatured a naive form of the argument in his *De libero arbitrio* (presented in quotes in the text):

> "If God has foreknowledge of my future will, then I am necessitated to will what He has foreknown, since nothing can happen differently than God has foreknown it. But if I am necessitated, we must admit that I no longer will freely, but of necessity." [III:3(8)]

While Augustine called this argument "sheer folly," it has nonetheless held sway over many people through the centuries (see Zabzebski, 2002, for a review of historical and contemporary views on divine foreknowledge and free will). PFW provides a naturalistic account of such sway, even under the assumption that God does not exist.

Let us suppose that Barrett (2000) is correct in asserting that widespread belief in God is a spandrel; the yield of a "Hyperactive Agent Detection Device" which conferred a fitness advantage upon our evolutionary ancestors. If such an account is correct, we should expect people to feel toward God the same sorts of feelings they feel toward other agents, including predators. As such, we might expect Mary to worry about God knowing what she is about to do, *whether or not she explicitly believes in God.* For, if belief in God is an evolved disposition, then its attendant worries and comforts ought

to be difficult to shake, as are the appeals of fat, sugar, and scantily clad conspecifics of the opposite sex. The nonexistence of God, though, means that Mary's PFW is safe from divine foreknowledge, worry though she may.

Harry Frankfurt (1969) introduced a controller who would force an agent to perform some action if and only if the agent was not going to do that action anyway.[2] Frankfurt never suggested that such controllers actually exist, and in fact allowed that the controller could be replaced by a machine or even natural forces (p. 836n). In contrast, PFW, defined as the ability to confound predators, depends critically on the sophistication of the actual predators in an agent's environment. For Mary, the hypothetical possibility of Frankfurt controllers is irrelevant; the important thing is their empirical absence. Thus, an agent worried about Frankfurt controllers should want PFW, in the same way that those who have traditionally worried about divine foreknowledge have wanted free will.

4 Moral Responsibility

Kant held that moral responsibility requires *transcendental freedom*: "Without this freedom...no moral law and no moral imputation are possible" (Kant, 1956/1788, p. 227). Parfit (2011) argues that Kant is confusing determinism with fatalism, and that morality requires only the sort of freedom which uses 'could' in the hypothetical, motivational sense (pp. 262-3). Here, as through-

[2]His point was that moral responsibility does not seem to require alternative possibilities.

out his book, Parfit conflates moral responsibility with moral accountability, the difference being that accountability requires interaction with other agents (Pereboom, 2001). This is understandable, since without recourse to a supreme arbiter (i.e. God), moral accountability is not practically different from moral responsibility. Thence, one of the primary questions facing an atheist moral philosopher like Parfit is the extent to which we can legitimately hold others accountable for their actions, where holding others accountable is the same as holding them responsible.

The sorts of actions for which we wish to hold others responsible tend to be those which harm us or others, particularly when they benefit the actor; i.e. *pure coercive transfers* in Judge Posner's typology (Posner, 1985). The victim, or a third-party observer, of a pure coercive transfer may want to hold the perpetrator responsible, and experience shows that it is easier to hold others responsible for acts which we do not anticipate. For our purposes, anyone who would hold Mary responsible for her actions is a predator. While PFW may help Mary by allowing her to complete a pure coercive transfer without being thwarted by such predators, the unpredictability it entails also makes it easier for them to hold her responsible for it.

Suppose Mary's action A_t^* at time t is a pure coercive transfer. If Mary has PFW, then from the point of view of anyone but God, she did have genuine alternative possibilities available to her at some time t: these were the elements, or potential elements, of S_t. As far as anyone can tell, she 'could have done otherwise' than to select and perform A_t^*. Her selection of

a particular action may have been random, but it was her preferences which posed it for consideration and her randomness which ultimately selected it.

Many libertarians in the free will debate agree with Bob Kane's assertion that free will requires 'ultimate responsibility,' which entails alternative possibilities (Kane, 1996, p. 33). In Kane's libertarianism, free will is exerted in self-forming actions (SFAs) by an agent confronted with multiple possibilities for which she has good reasons (Kane, 1996, p. 75). In practice, choosing A_t^* from a non-trivial S_t could meet all the criteria for an SFA. Mary had reasons, which were as much hers as anything, for each element of S_t. Nothing in the definition of PFW precludes the selection of a particular A_t^* from changing Mary's preferences or otherwise biasing her future behavior. Whether or not the world is deterministic, if the criteria for PFW are met then there is nobody around to notice the difference. Thus, PFW offers libertarians like Kane something akin to what they ask of free will.

5 Determinism and Indeterminism

Protean free will turns out to be possible whether or not determinism holds, but differently so in each case.

Suppose the world is not deterministic. Suppose Mary's brain is occasionally subject to 'Lucretian swerves,' ripples in the fabric of causation which cause some of her actions to be genuinely random. If the randomness is subject to Mary's control, then this is sufficient for PFW: Whenever the need

arises, she can define the set S_t, and then use her randomizing device to select an A_t^* at the very moment she is about to act. Since the selection is both truly random and made at the moment of action, it would be theoretically impossible for a predator to learn A_t^* at any time prior to t. Furthermore, as A_t^* is selected from S_t at time t, A_t^* will always be in S_t at time t. Therefore, this strong form of indeterminism is sufficient for PFW. However, we have no reason to believe that the world or the brain actually works this way.

Now, suppose the world is not deterministic, and Mary's brain is subject to occasional Lucretian swerves, but Mary has no control over when they happen or exactly what effects they have. This might be the case if the physical activity in Mary's brain were affected by quantum events. Therefore, her act of defining the set S_t is sometimes subject to random errors, as is her selection of an A_t^* to perform. If the swerves were frequent and severe, they would render Mary's brain nearly useless. At the other extreme, if the swerves were very subtle, their effects could be eliminated by rate coding (cf. London et al., 2010). This would leave Mary's brain effectively deterministic.

So, suppose Mary's brain is subject to some degree of purely random noise which does not debilitatingly interfere with her mental functioning, but which makes her behavior somewhat stochastic. Specifically, suppose that Mary has at least a minimal understanding of the sorts of effects that randomness has on her behavior. Now Mary can deal with this uncertainty in the same way that she deals with uncertainty about events outside of her body, whatever that way happens to be. Thus, this level of randomness does

not impede her ability to define S_t. If, at time t, a random event causes her to select some action B_t which is not in S_t, that is unfortunate. However, the fact that she did not in fact select an action in S_t does not mean that she did not have the ability to do so at time t. There is no reason to think she did not have this ability. A random event just caused her not to do so.[3]

Most importantly, at this level of impact, the randomness in her brain may have strategic value to Mary. First of all, the nature of the randomness in her brain and the steps that she takes to deal with it may be private information, in which case predators would be unable to know even the contents of S_t with certainty. Even if the nature of her neural randomness were common knowledge, its regular effects would make it impossible for a predator to know with certainty beforehand which action A_t^* Mary would perform at time t. A predator could still, in theory, predict Mary's action A_t^* with sufficient confidence that it could exploit her, and her chosen action would thereby be eliminated from S_t on a regular basis. Thus, PFW would be possible, but not trivial, in such a world.

On the other hand, suppose the world is deterministic. If all agents have bounded capacities of perception and reasoning, then determinism changes nothing. To all agents, including predators, the world still appears indeterministic. For Mary and PFW, the relevant level of apparent indeterminism depends on both the complexity of the causal relationships underlying her

[3]If the randomness is replaced with pseudorandomness, this argument still works, but is now more controversial; see Dennett, 2003, p. 81. The pseudorandom version is extraneous to the problem at hand.

behavior and the sophistication of the predators. If the predators are relatively sophisticated compared to the complexity of Mary's behavior, she could be deterministic enough for them to exploit. Thus, for example, while the protean flight path of the fly protects it from hungry birds, it is no defense against a human capable of building a fly trap. Therefore, it is possible that predators could preclude Mary's PFW. If Mary's world turns out to be the Matrix, of the eponymous 1999 film, or the organic supercomputer constructed by pandimensional beings, of Douglas Adams' *Hitchhiker* novels, then she does not, and could not, possess PFW.

The interesting case arises when Mary is among the most sophisticated agents in her environment, and so is of similar sophistication to some of her predators. Now Mary does not have to worry about superpredators, but she cannot rely too much on private information, either. This situation in particular calls for protean behavior. While Mary's ultimate selection A_t^* may be totally determined by prior circumstances, all that matters is that it not be determined or predicted by anything perceptible to her predators. It may be sufficient, therefore, to select A_t^* from S_t by some pseudorandom process. The process must take place in finite time, say over the interval $\tau := (t - \epsilon, t)$, prior to the moment of action. Mary's challenge, whether or not determinism holds, is to ensure that any 'noise' generated during τ by her choice process be uncorrelated with A_t^*.

6 Cloaking the Decision Process

I have shown that PFW is possible in both deterministic and indeterministic worlds, and that in either case it requires a decision process which does not hint at its outcome through the noise it generates. I will now consider the interval τ during which this noise poses a meaningful risk. This is the period just prior to an action during which that action has either been determined or has been distinguished as significantly likely. It exists because Mary is a human being, making decisions with a physical brain, which requires time and space to make choices and plan motor actions. Protean free will turns entirely on Mary's ability to 'cloak' the process during τ so that nothing perceptible to predators during this time is correlated with A_t^*.

A population of individuals all trying to conceal their behavior from each other, while trying to discern patterns in the behavior of others, will undergo an intraspecific symmetric arms race (cf. Dawkins and Krebs, 1979). In every generation, those individuals who are best at discerning the behavior of others, as well as those with the best concealment strategies, will have a selective advantage. As a result, the individuals who evolve over time will have sharp intuitive psychology and very good poker faces. The former will serve them when they are predators; the latter is what Mary needs for PFW.

7 The Libet Experiments

The chief law governing health insurance administration and the related privacy concerns in the United States, known as HIPAA, contains a 'need to know' provision for information about patients' health and healthcare. Within the massive bureaucracy of a contemporary American hospital, the only things a given staff member *should* know about a patient are those things which that staff member *needs* to know to do her job. This dictum exerts a profound influence upon the healthcare industry, in the way that documents are written, the way software is designed, and the way administrative workers do their jobs. Could Mary's brain implement a similar policy?

In his experiments in the 1980's, Benjamin Libet found that the 'readiness potential,' a scalp potential visible on an EEG prior to a motor action, preceded conscious awareness of certain actions by about 200 milliseconds (Libet et al., 1983). Libet himself claimed that his findings impose constraints on free will (Libet, 1999). However, his findings could also indicate a need-to-know policy operating in the brain. Indeed, the readiness potential not only poses no threat to PFW, but actually supports it. Suppose Mary has a demonstrated readiness potential.

Whether or not consciousness has privileged access to preferences, the readiness potential need not say anything about the process of approximating S_t. Once this set is defined, Mary is genuinely indifferent between its various elements. If she could tell you, at some point after she had defined S_t but

before she had selected A_t^*, what she most wanted to do at time t, she would say, "do one of the things in S_t without anybody else knowing what it's going to be." This is her choice, as much as it might be her choice to do some specific action B_r at time r. For her to be free to make this choice, she does not need to consciously know ahead of time which A_t^* she will actually end up performing.

Again, Mary is human, and things in her brain take time. Even after selecting an action, it takes time to translate that imagined action into a motor plan and enact it. Everything in her brain is connected, and these processes will generate some amount of noise. However, she should have evolved to keep this noise to a minimum. This will likely include keeping it away, as much as possible, from the great amplifier of consciousness. For Mary, the Libet experiments show a crafty tool nature has evolved to guard her PFW.

8 Is PFW Compatibilist?

Protean free will is broadly compatibilist in that in is compatible with both determinism and indeterminism. It is more rarefied than the 'commonsense' semicompatibilism of John Martin Fischer (Fischer, 2006). Nonetheless, PFW meets his various criteria for free will, which all center around 'our phenomenology as agents' (Fischer, 2007, p. 72). For Mary, the sensation of defining S_t and selecting A_t^* will be some form of the familiar process of

considering alternatives and making a choice. To her conscious experience, her ultimate choice may not seem random. This may make it seem all the more 'hers,' helping her to take responsibility for it and to learn from her actions. Thus, a protean-free action will feel as free as anything.

For the compatibilist Daniel Dennett, free will is tied not only to moral responsibility, but also to reasons (Dennett, 2003, esp. Ch. 9). For Dennett, a free agent ought to be able to offer up reasons for any past action when called upon to do so. Were Dennett's Inquisitor to observe Mary doing A_t^*, and ask her why she had done so, Mary would have no shortage of reasons to offer. How, though, is the Inquisitor to differentiate between reasons for the inclusion of A_t^* in S_t, and reasons for its ultimate selection? He may not be able to tell! Thus, the sort of accountability to others which Dennett emphasizes need not undermine Mary's PFW. The Inquisitor does not ask Mary what other actions she had equally good reasons for at time t, and even if he did, Mary might not be able to answer. Cloaking the decision process may mean that some possibilities, which were real possibilities for her at some point prior to t, never entered her conscious awareness. Even if she were to tell the Inquisitor everything she knows about how her mind works and what she plans to do in the future, she might still then beat him in Rock-Paper-Scissors.

PFW could also be a consolation prize for incompatibilists. For the hard incompatibilist Derk Pereboom, moral responsibility is the point of free will, which requires 'robust' alternative possibilities (Pereboom, 2001, p. 26).

As described above, PFW can satisfy the requirements of such responsibility, provided that the people (or whatever) who hold Mary responsible for her actions may be viewed as predators. This explains why Pereboom can make a perfectly coherent argument for hard incompatibilism to a Christain audience, as in Pereboom (2005), while his God-free presentation of hard incompatibilism in *Four Views* (Pereboom, 2007) seems to be a quagmire of question begging.

Likewise, the incompatibilist Peter van Inwagen's *Consequence Argument* (van Inwagen, 1983, p. 16) requires a God's-eye-view to do any work. For even if our actions are uniquely and entirely caused by the past and the laws of nature, no mortal being who is affected by our actions could possibly care that this is the case. Van Inwagen literally invokes a God's-eye-view when he argues for the incompatibility of free will and indeterminism:

> Now let us suppose that God a thousand times caused the universe to revert to exactly the state it was in at t_1 (and let us suppose that we are somehow suitably placed, metaphysically speaking, to observe the whole sequence of "replays"). (van Inwagen, 2000, p.14)

Should van Inwagen desire something akin to free will which requires no metaphysical heroics, he might find PFW a suitable alternative.

9 Concluding Thoughts

I have presented a notion of free will with a different emphasis from that of most philosophers who have written on the subject. Protean free will is not mutually exclusive, nor even necessarily different, from other notions of free will. My original intent was to look directly at the role which other agents play in the definition of free will, a role which I found underrepresented in the current literature. In order to make an evolutionary argument, I focused on agents with competing interests, which I called predators. Inspired by the work of the evolutionary psychologist Geoffrey Miller (1997), I chose to focus on protean behavior as a defense against such predators. The resulting form of free will turns out to be something which most people probably possess, but not trivially so, and which satisfies many of the requirements posited by philosophers of various stripes.

Furthermore, the importance of PFW in strategic interactions suggests an evolutionary basis for the grip which issues of free will have held on so many minds. Centuries ago, believers in a highly anthropomorphized God thought of God the way they thought of other humans, which included worrying about God knowing what they were going to next. While few contemporary theologians take the issue seriously, the same worries have been taken up by quite a number of contemporary philosophers.

Thus, many of the philosophical problems currently associated with free will arise when philosophers reject (or ignore) the notion of a personal God,

but retain some of the attendant baggage. Without God, others foreseeing and thwarting one's actions is a practical problem, not a metaphysical one. Without God, moral responsibility is indistinguishable from moral accountability, and there is no coherent sense in which this can require libertarian free will. Without God, free will offers little to nothing that protean free will does not.

References

Augustine of Hippo (1968/ca. 391): "The Free Choice of the Will," in *The Fathers of the Church*, Washington, D.C.: The Catholic University of American Press, vol. 59, trans. Russell, Robert P.

Barrett, J. (2000): "Exploring the natural foundations of religion," *Trends in Cognitive Sciences*, 4, 29–34.

Chance, M. and W. Russell (1959): "Protean displays: a form of allaesthetic behaviour," *Proceedings of the Zoological Society of London*, 132, 65–70.

Dawkins, R. and J. R. Krebs (1979): "Arms Races between and within Species," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 205, 489–511.

Dennett, D. C. (2003): *Freedom Evolves*, Penguin.

——— (2006): *Breaking the Spell*, Viking.

Descartes, R. (2008/1641): *Meditations of First Philosophy, with Selections from the Objections and Replies*, New York: Oxford University Press, trans. Moriarty, Michael.

Fischer, J. M. (2006): *My Way: Essays on Moral Responsibility*, New York: Oxford University Press.

——— (2007): "Compatibilism," in *Four Views on Free Will*, Blackwell.

Frankfurt, H. G. (1969): "Alternate Possibilities and Moral Responsibility," *The Journal of Philosophy*, 66, 829–839.

Kane, R. (1996): *The Significance of Free Will*, New York: Oxford University Press.

Kant, I. (1956/1788): *Critique of Practical Reason*, New York: Liberal Arts Press, trans. Beck, Lewis White.

Lewis, C. S. (1940): *The Problem of Pain*.

Libet, B. (1999): "Do we have free will?" *Journal of Consciousness Studies*, 6, 47–57.

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.

Locke, J. (1985/1689): *An Essay Concerning Human Understanding*, Oxford: Clarendon Press.

London, M., A. Roth, L. Beeren, M. Häusser, and P. Latham (2010): "Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex," *Nature*, 466, 123–127.

Miller, G. F. (1997): "Protean primates: The evolution of adaptive unpredictability in competition and courtship," in *Machiavellian Intelligence II*, ed. by A. Whiten and R. W. Byrne, Cambridge: Cambridge University Press.

Parfit, D. (2011): *On What Matters*, Oxford: Oxford University Press.

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

——— (2005): "Free will, evil, and divine providence," in *God and the Ethics of Belief*, ed. by A. Chignell and A. Dole, Cambridge: Cambridge University Press, 77–98.

——— (2007): "Hard Incompatibilism," in *Four Views on Free Will*, Blackwell.

Posner, R. A. (1985): "An Economic Theory of the Criminal Law," *Columbia Law Review*, 85, 1193–1231.

Samuelson, P. A. (1947): *Foundations of Economic Analysis*, Cambridge: Harvard University Press.

van Inwagen, P. (1983): *An Essay on Free Will*, Oxford: Clarendon Press Oxford.

——— (2000): "Free will remains a mystery: The eighth Philosophical Perspectives lecture," *Nos*, 34, pp. 1–19.

von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press.

Zabzebski, L. T. (2002): "Recent Work on Divine Foreknowledge and Free Will," in *The Oxford Handbook of Free Will*, ed. by R. Kane, New York: Oxford University Press.