Fisher, Justin. (forthcoming) "What Can Game Theory Tell us About Humans?" to appear in *Evolution, Games and God*. Sarah Coakley and Martin Nowak, editors. Harvard University Press.

What Can Game Theory Tell Us About Humans?

Justin C. Fisher – Southern Methodist University

The preceding chapters have offered various game theoretic models of cooperation, and have suggested that such models can help to shed light on patterns of human behavior. One natural response to these models is to grant that they might help us to understand computers and simple animals, but to be skeptical about how much these models might tell us about humans. In this chapter, I examine three potential reasons for such skepticism. Each of these potential reasons involves a feature that many people suppose sets humans apart from simpler creatures:

(1) **Mind-Body Dualism.** Many people suppose human minds to involve some sort of non-physical thinking substance, distinct from the physical substances that compose our brains and bodies. Since science usually deals only with physical things, one might worry that standard scientific approaches, including game theory, will have an especially difficult time accounting for non-physical minds.

(2) **Free Will.** Many people suppose that human actions are caused just by ourselves, and not by the various factors which led us to be the way we are. Since science usually deals only with things that are fully under external causal influence, one might worry that it will have a difficult time accounting for freely chosen actions.

(3) **Complexity.** Many people are struck by the rich capacity of humans to behave differently across a wide variety of circumstances. Since existing game theoretic models include only a handful of parameters, one might worry that these models will be ill-suited to account for the rich complexity of human behavior.

At first blush, these three concerns seem logically independent of each other: one can imagine possible views that accept any combination of these and reject the rest.[1] Nevertheless, many people who are moved by some of these concerns are also moved by the others. Each of these concerns might be motivated by attending to the rich diversity of human behavior, and especially to our capacity to behave in ways that are creative, inspired, and/or unpredictable. This rich diversity suggests (3) that human psychology must be very complex. This complexity might seem to recommend thinking (1) that the human mind contains something non-physical in addition to the few hundred billion neurons and the quadrillion synapses that compose the human brain. And if our complex behavior is to be unpredictable not just in practice, but in principle, then this might seem to recommend thinking (2) that it is produced by some special sort of "agent causation" which comes from outside the web of physical causal relations.

[1] I say "at first blush" because it may be questionable whether certain views of libertarian free will are actually coherent at all. If libertarianism contains logical contradictions within itself, then it certainly won't be logically compatible with any other views.

My own view is that we currently have no reason to expect that the complexity or practical unpredictability of human behavior implies that human minds contain anything beyond the many billions of neurons and synapses that compose our brains. So, of these concerns, I am personally gripped only by (3). However, I recognize that (1) and (2) will also be gripping to many readers, especially adherents to popular religious traditions. This chapter takes (1), (2) and (3) seriously, and addresses the question of whether someone gripped by these should therefore have significant worries about the game theoretic approach to understanding human behavior outlined by other authors in this book. My conclusion will be cautiously optimistic: game theory is compatible with all plausible positions regarding dualism and free will, and, while there is much room for game theoretic models to improve with respect to the complexities of human cognition, these are improvements that we may expect game theorists eventually to make.

Mind-Body Dualism.

There are various potential motivations for dualism. These include the other two concerns mentioned above: one might doubt whether purely physical substances could produce the rich complexity of human behavior, and/or whether they could do this in a way that is appropriately free from outside causal influences [***link to Zimmerman, this volume?]. In addition, many people have doubted whether purely physical substances would be capable of the conscious experiences that humans have.[2] Several religious commitments might also motivate dualism. If the mind is to survive the destruction of the body (to go on to heaven or hell or reincarnation or wherever), then mind apparently must be distinct from body.[3] And if embryos are to be full-fledged persons and/or if they are to have 'original sin', this might militate in favor of supposing that they have non-physical minds even before they have brains.

I won't attempt to evaluate these motivations here.[4] Instead, my goal in this section is to ask whether dualism, *if it were true*, would pose a threat to the game-theoretic modeling undertaken in this book.

A great deal of scientific work is geared towards understanding purely physical systems. One might worry about the potential to extend scientific models to non-physical systems,

---

[2] For a good survey, see Chalmers 2003.

[3] Some elaborate accounts of the afterlife are compatible with the view that living humans are purely physical beings. For example, one might hold that, at the moment of death, human bodies are transported away to the afterlife and replaced with look-alike corpses (van Inwagen 1978). Or, some physicalists hold that there is a potential for uploading our minds into computers (Bostrom 2004), and one might imagine that a deity has arranged it so that something similar already happens at the time of bodily death.

[4] The editors have encouraged me to explain why I myself am not attracted by dualism. My strongest reason is Occam's razor. We know that brains exist, that they are immensely complex, and that many physical changes in brains cause changes in behavior. Given all this, the default hypothesis is that our brains are in sole control of our behavior. We would need a compelling reason to posit some further thing helping to control of our behavior. As I'll note below, I don't accept views of free will that would motivate dualism. Regarding consciousness, I'm not so confident of my introspective abilities that I would adopt dualism merely on the basis of the fact that my conscious experiences don't seem physical (c.f., Dennett 1991). And I hold no religious beliefs that require dualism.

as the dualist takes human minds to be. For example, Princess Elisabeth of Bohemia famously argued that there was no way of understanding how a material brain and an immaterial mind could interact (Blom 1978). Such worries would be especially pressing against ambitious scientific research programs that hope eventually to show how all aspects of the world are composed of the same basic elements governed by the same basic laws.

In response to these concerns, I will argue first that there is no principled reason why there couldn't be a science of non-physical substances, and second that game theoretic models, in particular, are especially amenable to the possibility that human minds will turn out to be non-physical.

Let's begin with a useful analogy. Consider someone who has been completely immersed in a multiplayer online game since birth – virtual reality is the only reality she knows. Through clever experimentation, she might eventually discover all the laws governing the 'physics engine' of her virtual world. However, she would also discover that the characters in the game sometimes behave in ways that are consistent and reasonable, but completely unpredictable from within the physics engine. A reasonable conclusion would be that some things outside the game (namely the human players) are controlling the characters within the game. This might be the end of the road for developing a 'physics' of this game, but it wouldn't be the end of *science*, for our clever player could set up various experimental situations involving characters within the game (including herself), and see how the outside controllers for those characters react. Depending upon the richness of the game interface, our clever scientist might glean a great deal about human psychology, and perhaps even about the more general laws of the world outside her game.[5]

The dualist holds that our own predicament is very much like that of the unwitting player of a virtual reality game. The dualist holds that when we completely decipher the 'physics engine' of our universe, some events in our brains will be left inexplicable. Given such findings, it would be reasonable to conclude that, like the characters of the game, our own brains and bodies are controlled by something outside of them. And just as our clever game player could go on to test scientific hypotheses about what was controlling the characters in her game, we might someday go on to test scientific hypotheses about mental substances controlling our brains and bodies. Depending upon the richness of the mind-brain interface, we might glean a great deal about the structure of our minds, and perhaps even about more general laws of the world our minds inhabit.

So, there is no principled opposition between dualism and a general science of human cognition. There may, however, be a tension between dualism and particular scientific approaches to human cognition. E.g., dualism is incompatible with neuroscientific approaches that aim to explain all cognition via the physical characteristics of neurons and other brains structures [***link to Ned Hall's Crass Materialist Atheist Reductionists, this volume???]. Someday we'll need to choose between dualism and neuroscience. But

[5] Notice that, in this case, Princess Elisabeth's worry (how could mind and body possibly interact?) completely evaporates.

this is a book in game theory, not neuroscience, so we can leave that choice to another day. Let us now ask whether *game theory* makes assumptions that are incompatible with dualism.

Most game-theoretic models do make many assumptions about the players that play the games. E.g., most models presume that the players have certain preferences, that they are fairly well-informed about the structure of the game, and that they will be quite rational in choosing strategies that would do well to satisfy their preferences in the circumstances they believe themselves to be in.

However, these assumptions are entirely neutral about what sorts of substances the various players of the games are made out of. So long as the players have the relevant beliefs and preferences, and so long as they choose their strategies in the specified ways, game theory will apply to them, regardless of whether the players are made from flesh or silicon or ghostly ectoplasm. Unlike neuroscience, game theory considers human cognition at a level of abstraction which stakes no particular claims regarding what exactly humans are made of, and hence game theory is fully amenable to the possibility that dualism might turn out to be true.

Free Will.

A distinction is commonly drawn between 'compatibilist' and 'libertarian' understandings of free will. According to the compatibilist, the fact that we sometimes act freely is compatible with the possibility that our world is deterministic – that given the state of the world (including any non-physical substances it might contain) at one time, the laws of nature fully determine how all subsequent events (including all human actions) will proceed. E.g., a compatibilist might hold that, so long as my actions are produced by healthy deliberative processes they will count as "free actions", even if those actions (and the deliberative processes that produced them) were causally determined by my genes and my upbringing. Since compatibilism views our actions as being fully embedded in the causal structure of the world, there is no special tension between compatibilism and scientific approaches like game theory.

In contrast to compatibilism, libertarianism holds that human actions are produced by a special sort of 'agent causation' in a way that makes them not be fully caused by preceding events. Since the libertarian thinks human actions are quite different from ordinary physical events, she might worry that scientific approaches, like game theory, that are well suited for explaining other events wouldn't work well at explaining human actions. Our goal in this section will be to explore these concerns.

There are several potential motivations for libertarianism. Human behavior seems spontaneous and unpredictable, and it would be very disturbing to learn that, no matter how hard we try, the current state of the world fully determines what all our future actions will be. There are also strong intuitions that we do act freely, and that our actions wouldn't be free if some state of affairs outside our control – e.g., the complete state of the world before we were born – was entirely sufficient to cause our actions to occur (van

Inwagen 1983). If we embrace both these intuitions, then we must suppose that our actions are somehow free from antecedent causes outside our control. Various religious commitments might also motivate an acceptance of libertarianism. If a world with libertarian free will would somehow be better than a world lacking it, then an omni-perfect creator would need to have included it (Murray 1993). Libertarianism may also be needed for the 'free will defense' to the 'problem of evil': the defense that blames all suffering on human free choice and thereby absolves the creator of responsibility – this would be hard to do if the creator's act of creation was itself sufficient to cause all these choices and the ensuing suffering. [***Link to Rota, this volume?]

As above, I will not attempt (at least not directly) to evaluate these motivations,[6] and will instead ask whether libertarianism, *if it were true*, would pose problems for the game theoretic models in this book.

Many game theoretic models are deterministic – they presume that whenever you put an agent with certain preferences into a certain sort of circumstance, the agent will definitely behave in a certain way. (E.g., in some models agents will definitely do the rational thing; in others they will definitely imitate their most successful neighbor.) This determinism is in apparent tension with libertarian free will, which allows that agents may freely choose to do any number of things in a given circumstance.

Game theory does have one tool for accommodating uncertainty regarding how people will behave, namely allowing that agents might behave stochastically, choosing different strategies with different probabilities. However, determinately employing a fixed probability distribution would likely strike libertarians as being no more 'free' than was determinately employing a single fixed strategy (Mele 2006). If libertarians insist that there is significantly more to free action than just randomly choosing actions in accordance with predetermined probability distributions, then there will be tension between libertarianism and game theory.

In response to these concerns, I will argue (a) that game theory actually has fairly minimal commitments regarding the causal and probabilistic relations between human circumstances and human choices, and (b) that any plausible version of libertarianism must be compatible with these minimal commitments.

---

[6] The editors have encouraged me to explain why I myself am not attracted to libertarianism. My main reasons for this are, first, that libertarianism strikes me as being either incoherent or else too bizarre to be plausible, and, second, that the intuitive evidence that supposedly favors libertarianism instead seems, on reflection, to support compatibilism. Why would libertarianism be incoherent or bizarre? Because it is difficult to see how my actions could be sensitive to my antecedent knowledge, desires, and plans, and yet be free of antecedent causes. Even if I pretend it makes sense to posit some special sort of "agent causation" here, Occam's razor cautions against positing such things without good reason. Why don't I see a good reason? My actions don't seem to me to be uncaused by antecedent factors, and even if they did seem that way, I doubt I would trust such seemings. As for intuitions that free action and moral responsibility are incompatible with our actions' having antecedent causes, it seems to me that our systems of education, reward and punishment presume that things we do *can* cause changes in what people will freely choose, and hence I think that, on reflection, we are intuitively committed to compatibilism at least as strongly as we are to libertarianism. And I have no religious beliefs that require libertarianism.

One of game theory's commitments is that, if a model is to explain actual instances of human behavior, then the distribution of behaviors predicted by the model must match fairly closely the actual distribution of behavior in the world. For example, a model which says that 60% of people in a certain circumstance will make choice A can be a good explanation of people's actual choices only if approximately 60% of people in that circumstance make choice A. It is quite challenging to come up with models that actually do fit the complex diversity of human behavior, but that's a topic for the next section. For now we're concerned only with the question of whether doing so would be incompatible with libertarianism.

It is clear that there are facts about the probabilistic distributions of human choices. E.g., an exit poll of 10% of voters provides a very reliable estimate of how the other 90% will vote. One might take facts like this to suggest that there must be some sort of causal story, or at least some sort of probabilistic story, to be told about how all these voters make these choices; and one might worry whether such a story would be compatible with libertarianism. However, our present goal is not to make difficulties for libertarianism, so let's suppose that libertarianism can, somehow or other, accommodate these clear facts. But if libertarianism can accommodate such facts, as it apparently must, then we'll need to look elsewhere to find a tension between it and game theory.

Good scientific explanations do not just pick out interesting patterns in the distribution of various features in the world – they must also provide a guide to the causal structure of the world, and, in particular, they must provide a guide to intervening in that structure to bring about different results (Woodward 2003, Fisher 2006). So, a second commitment of game theory is that various factors that its best models take to determine agents' behavior must be factors that actually are causally relevant to human behavior. For example, many game theoretic models hold that a game's payoff structure helps to determine what players will do. If these models are to be explanatory, then it must be the case that changing the payoff structures of actual scenarios is a good way of bringing it about that agents will make different choices.

There are very strong reasons for thinking that changing payoff structures *can* change how people will act. The entire point of posting a reward for a lost pet is that doing so might cause someone to call in – and sometimes it does. One primary justification offered in favor of having a system of lawful punishments is the fact that (at least for many sorts of crimes) such systems deter people from behaving poorly. Once again, one might worry about whether libertarianism is compatible with these commonly accepted facts (Ayer 1954). But since our present aim is not to make trouble for libertarianism, let us suppose that, somehow or other, libertarianism can be made compatible with them. So we find that this second commitment of game theory is also compatible with any plausible version of libertarianism.

Does game theory have other commitments that *would* be in tension with libertarianism? I think the answer is no. Much as game theory was neutral with respect to dualism, game theory is quite neutral with respect to exactly what sorts of processes produce human action. I've noted that game theory does presume that these processes produce certain

distributions of behaviors in groups of people, and that they are sensitive to interventions on at least some external factors, like payoff structures. But, I've argued that if there are any plausible versions of libertarianism, they must be compatible with these presumptions. Insofar as libertarians can cook up a story compatible with these presumptions, game theory will be neutral regarding the question of whether we should accept that story.

Complexity.

We've seen that evolutionary game theory need not conflict with either dualism or (any plausible version of) libertarianism. Let's turn, then, to the third concern listed above – namely that game theoretic models won't be able to accommodate the complexity of human cognition.

Human choices seem to be predicated upon incredibly many factors. Our choices apparently depend upon fine-grained beliefs about our current circumstances, upon the relative strengths of our various desires, upon habits built through past learning experiences, and upon idiosyncratic personality traits and temporary moods. In contrast with all this complexity, existing game theoretic models usually involve only a handful of parameters, and indeed, game theorists often strive to "keep it simple" and not add unnecessary parameters to their models. This might raise worries that game theoretic models are just too simple to capture the sorts of complexity present in human behavior.

In response to this concern, I will first consider how far simple models can take us, and then move on to consider prospects for extending such models to accommodate more complexities.

Simplicity in an explanation isn't necessarily a bad thing. We may illustrate this point with an example from Hilary Putnam (1975). Suppose I have a square wooden peg with 2" sides, and a wooden board with a round hole 2" diameter. How shall we explain the fact that the square peg won't fit through the round hole?

One proposed explanation might laboriously enumerate all the possible orientations of the peg, and detail for each of these which parts of the peg would collide with which parts of the board surrounding the hole. This proposed explanation might convince us that the peg won't fit, but it won't afford us an intuitive understanding of why it won't fit, nor will it help us to approach similar cases.

An alternative explanation would point out that no rigid peg will be able to pass perpendicularly through a hole if it the peg contains two points in its cross-section that are further apart than the diameter of the hole. Since the two corners of our peg are more than 2" apart, it therefore can't fit through our 2" diameter hole. This alternative explanation makes no attempt to capture the full complexity of all the ways in which our peg might collide with our board. Instead it shows how our peg is just one instance of a general pattern involving many pegs and many holes. This general pattern allows us to predict which pegs can pass through which holes, and it gives us recipes for intervention,

telling us, for example, how much we'd need to whittle our peg to get it to fit through our hole.

For very many purposes, we want our explanations to be like this simple alternative explanation, highlighting simple patterns that are predictively and pragmatically useful. Once we see this general pattern, a detailed list of potential collisions is quite beside the point – these extra complexities add little, if anything, to an explanation in terms of the simple general pattern cited above. Good explanations are supposed to help us see the forest, not just catalog all the trees.

Similar considerations apply to game theoretic modeling. Game theoretic models clearly make no attempt to capture all the psychological complexities that play a role in producing human decisions. Instead, game theoretic models attempt to capture some small number of parameters (e.g., features of the payoff structure, or heuristics for imitating neighbors' strategies) and show how these parameters can be used not only to predict what humans will do, but also to give us recipes for intervention, telling us, for example, what sorts of incentives we would need to offer in order to get players to engage in cooperative ventures.

There is a lot to be said for an explanation that picks out a few highly relevant factors, and shows how a pattern of results depends upon those factors. It is quite challenging to come up with a model that captures a great deal of data using only a few parameters, especially parameters that give us a handle for intervening to bring about particular outcomes. Models like these are explanatorily valuable, often more so than models which use a large number of parameters to better fit the data.

But ultimately, we want both to see the forest and to know about all the particular trees. We want our explanations not just to highlight simple highly relevant patterns, but also to give us links to more detailed information, more accurate predictions, and more nuanced interventions. In Putnam's square peg example, an ideal explanation wouldn't *just* tell us that the corners of the square are too far apart to fit through the hole, but it would also offer us links to more detailed information, for example, about how rigid objects maintain their geometrical structure during collisions. Similarly, ideal explanations of human behavior won't *just* highlight the simple patterns that current game theoretic models highlight. Instead, they must also offer links to more detailed information, e.g., information about how humans manage to arrive at choices that are fairly rational.[7]

It is difficult to say how well game theory currently measures up in these regards. As a case in point, Dreber & Almenberg (this volume) describe a number of games that experimental economists have watched human subjects play. There are two general ways in which economists' predictions have differed from observed behavior of human subjects. One difference is that human subjects apparently care more about other players' payoffs than economists originally expected. Dreber & Almenberg note that this might be accommodated in game theoretic models either by presuming that players have a

[7] For an account regarding how explanations might offer such links, and regarding where these links should lead, see Fisher (2006, chapter 5).

different preference structure than was originally thought, or by presuming that subjects are choosing in something less than a fully rational manner. A second difference was that behavior varied significantly across human subjects. For game theory to accommodate this variation, it will likely need to introduce further parameters – perhaps different subjects have different beliefs about what consequences their actions would have, or perhaps they have different levels of other-regarding desires. Ultimately, one would hope that whatever tweaks game theorists make to their models could be empirically motivated by psychological evidence involving how humans make decisions. An honest assessment of this work must acknowledge that much behavior of human subjects isn't accounted for by current game theoretic models. So there is a great deal of room for improvement here, but also reason to hope that improvement will occur.

Critics of game theory will likely emphasize the gaps between the predictions of current game theoretic models and the wide variance in human behavior, and also the gaps between simple game theoretic presumptions about psychology and the complex psychological story that surely underlies human decision making. Such criticisms are valid in that they highlight ways in which our current best explanations fall short of the sorts of explanations we'd like to find. These criticisms are also valuable, in that they help to point the way toward future models that will be better integrated with our understanding of human psychology, and will make better predictions.

There are good grounds for optimism here. In the preceding sections, we considered worries about dualism and libertarianism that threatened to show that game theory was the *wrong kind* of theory for describing human cognition. In contrast, the present worry involves not a matter of *kind* but just a matter of *degree*: Can game theoretic models accommodate *enough* of the complexities of human cognition? This worry allows that game theory might at least be the *right kind* of theory for the task – it just insists that we need to find richer and more complex models of this sort.

This is good news for the game theorist. For if the bar is complexity, complexity is something that always can be added to game theoretic models. Even if current game theoretic models leave a great deal of room for improvement *vis a vis* the complexities of human psychology, these are improvements that game theorists can and surely will make.

References.

Ayer, A.J. (1954). "Freedom and Necessity". In *Philosophical Essays*, London: Macmillan.

Blom, J. 1978. *Descartes, his moral philosophy and psychology*. New York: New York University Press.

Bostrom, N. 2004. "The Future of Human Evolution." *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy. Palo Alto, California: Ria University Press, pp. 339-371.

Chalmers, D. (2003). Consciousness and Its Place in Nature. In S. Stich and F. Warfield, eds. *Blackwell Guide to the Philosophy of Mind*. pp. 102-142. Blackwell. Also online at http://consc.net/papers/nature.html

Dennett, D. (1991) *Consciousness Explained*. Boston: Little, Brown.

Fisher, J. (2006) *Pragmatic Conceptual Analysis*. University of Arizona, Ph.D. dissertation.

Mele, A. (2006) *Free Will and Luck*. New York: Oxford University Press.

Murray, M. (1993). "Coercion and the Hiddenness of God," *American Philosophical Quarterly* 30, 27-38.

Putnam, H. (1975). "Philosophy and Our Mental Life," in *Mind, Language and Reality, Philosophical Papers*, Vol. 2, ed. H. Putnam. Cambridge: Cambridge University Press.

Van Inwagen, P. (1978). "The possibility of resurrection," in P. Edwards (ed.) *Immortality* Amherst NY: Prometheus Books, 1997, pp. 242–246.

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

Woodward, J. (2003). *Making Things Happen: An Account of Causal Explanation*. Oxford: Oxford University Press.