5-17-2023

Intricacies of Agency: Rational Choice, Behavioral Economics, and Our Normative Commitments

Max Hendrix

Intricacies of Agency

Rational Choice, Behavioral Economics, and Our Normative Commitments

Max Hendrix

College of the Holy Cross
Fenwick Scholar 2023

Advised by: Karsten Stueber, Kolleen Rask, Mark Hallahan, and Daniel Schwab

Abstract

This project undertakes a philosophical analysis of the intricacies of agency found in rational choice theory - the mainstream economic theory that agents are fundamentally rational and utilize their rationality to identify and pursue their self-interest. Recent experimental evidence within and outside of economics has cast doubt on the psychological accuracy and predictive prowess of the theory, laying the foundation to discuss the strengths and limitations of the theory as well as the impacts that this paradigm of agency has on our society today. I argue that rational choice theory struggles as a holistic conception of agency both from an explanatory and a predictive standpoint because it fails to be psychologically accurate. Economics' push to become more precise using mathematical functions must be grounded in accurate conceptions of agency and society and returning to Adam Smith, an economist and philosopher, can serve as a bridge between the disciplines.

Introduction

I, for instance, would not be in the least surprised if all of a sudden, apropos of nothing, in the midst of general prosperity a gentlemen with an ignoble, or rather with a reactionary and ironical, countenance were to rise and, putting his arms akimbo, say to us all, "I say, gentlemen, hadn't we better kick over the whole show here and scatter rationalism to the winds, simply to send these logarithms to the devil, and to enable to live once more at our own sweet foolish will!"

Fyodor Dostoyevsky, *Notes from the Underground*

Most of the last 400 years have seen the study of economics intimately connected to philosophy and political organization. Philosophers and political thinkers saw commerce as one important element that must be considered alongside concerns of morality, politics and justice when deciding ways to govern and live one's life. Historical scholars such as Adam Smith, Thomas Hobbes, and David Hume were all well recognized philosophers who made contributions to political theory, philosophy, and economics in their discussions of how we should organize our society. None of them were economists in the sense of the word used today, but they all laid important philosophical foundations for the way that social scientists would eventually come to view human action and decision making in the 20th and 21st century. Today, economists serve as political advisors to the highest government offices with an expansive grip on the way that our states and countries are governed. Within the United States, impact to the economy is always one of the foremost considerations for any political plan that is proposed, and thus economists have a large role to play in deciding how we allocate our resources. This places a high burden on economics to be accurate about the wide range of motivations that affect individuals in order to properly engineer our economy in a way that upholds our commitments to each other. In the recent decades, economics as a discipline has distanced itself from the folk psychological theories of its historical proponents and sought to cast itself more as a science with

powerful predictive capabilities but losing the complex elements of human beings such as their commitments and beliefs.

The ongoing discussion of the role of economics as a social science within our society hinges on the fundamental questions of human nature and the constitution of our agency. Agency is a term used to describe beings who act for reasons. For the purpose of this paper, I will be primarily concerned with human beings when I refer to agents. Our mental states, such as our beliefs and desires can be understood as reasons that can act as causes for our actions. Since an agency is so dependent on reason-based action, all theories of decision making (which is the process by which we go about selecting the motivation that ultimately manifests itself as an action) are inherently connected to the discussion of agency. This is not to say that *all action* is related to agency, but that all agents are defined by their actions.

Philosophy, economics, and psychology all focus on agency from different perspectives but collectively seek to understand the constitution of human agency. I will focus on two primary philosophical camps into which paradigms of human agency fall. Both sides understand and recognize that there is some mixture of rationality, self-interest, and sympathy within agents, but each views decision making as being motivated by different sources. Proponents of rational agency, such as Thomas Hobbes, view agents as primarily self-interested (cooperation and altruism can exist) and guided by their instrumental rationality to select the best possible action for themselves. In contrast, moral sentimentalists emphasize that although humans are rational, their agency is socially constituted, and they are guided to action by their emotions and sympathy (their passions) for others in addition to reason.

Adam Smith and the moral sentimentalist philosophers argue that decision-making is always a product of the environment that we live in because our desires are influenced by the

judgement of other people. Smith saw that individuals cultivate desires, emotions, and beliefs in relation to others' approval and disapproval of our actions. Our desires are not independent of the rest of the world or formed only in our own mind but are contextually dependent. He believed that as a society develops, morality will develop right alongside which helps to constrain an individual's self-interest. Smith articulated this position mostly in his second most famous book, *The Theory of Moral Sentiments* (TMS). However, the work that he is most famous for is his magnum opus, *The Wealth of Nations* (WN). Economists regard him to be the father of classical economics, and they read his view of agency much differently. They understand Smith as telling us that society will function completely well even when each individual acts in their own self-interest[1]. This reading argues that Smith believes that even though agents are guided by their own self-interest, the power of the market system guides us to an equilibrium outcome that is good for society. As economics continues to remain important within individuals' daily lives through government decisions and actions, it is ever more imperative that we reflect and explore where its view of agency fits.

Within economics, I contend that the mainstream view of agency is described through rational choice theory[2], which is made up of two main economic tools/models: expected utility theory and game theory. Rational choice theory assumes that human agents are rational beings who are guided by their desires to maximize their own personal payoffs (These payoffs are not necessarily merely monetary). Typically, self-interest is the first consideration for the best interest of an agent, but this interest can involve altruistic motivations. This theory sets the groundwork for *laissez-faire* economics which relies on the assumption that decision making by

[1] You can see this reading of Adam Smith's writing in any introductory economic textbooks such as Mankiw 2018.
[2] Rational choice 'theory' is not a theory like the theory of evolution, rather it is a broad collection of assumptions, models and tools that are used in economics to predict agent's actions. You can think of it more as a paradigm of agency than a theory in the traditional sense.

agents is fundamentally rational and governed by the goal of maximizing their own economic payoffs. The view of agency in economics influences the way that we structure our societies by impacting the institutions that encourage social cooperation and prevent individuals from taking advantage of others. Our conception of agency influences many facets of our life, so having an accurate picture is crucial for us to make well-informed decisions. The picture of agency that was portrayed in rational choice theory is intriguing because it seems to be focused on gaining mathematical precision at the cost of implausible assumptions about agency; a big change in comparison to the rich image of an agent described by Smith. Recent empirical evidence from behavioral economics and psychology has cast doubt on the prowess of rational choice theory, laying the foundation for a philosophical analysis of the importance of accuracy in paradigms of agency both in terms of prediction and explanation. This paper will address the potential strengths and complications of adopting these psychologically inaccurate assumptions and explore the challenges faced by rational choice theory while explaining the usefulness and power of the theory that has dominated the field of economics for almost a century.

It has become quite clear that there is significant evidence that contradicts the traditional conception of agency in economic theory. Game theory researchers, such as Colin Camerer (Camerer 2003) and Cristina Bicchieri (Bicchieri 2018) among many others, have conducted laboratory studies that ask participants to decide how to split money between themselves and someone else, with varying constraints and options given to them. These studies, in what has become the foundation of behavioral economics, overwhelmingly returned results in which participants did not merely maximize their own returns but actually showed concern for others with no benefit to themselves (pure altruism) and commitment to normative principles (fairness, justice, friendship etc.). Kahneman and Tversky show how important framing of a choice is

when individuals are faced with decisions (Kahneman & Tversky 1979); Herbert Simon casts doubt on economic decision makers' ability to even find the maximizing alternative let alone select it from all of our alternatives (Simon 1955); and Daniel Batson argues expertly for the existence of true altruistic motivations within human beings (Batson 2010).

Although the theory fails to be psychologically accurate, I argue that it can still be predictively useful in specific, constrained situations – such as building rules for an auction. Furthermore, I will argue for the importance of achieving more psychologically realistic and socially embedded theories of agency and how Adam Smith may be able to serve as the bridge between rational choice theory and conceptions of agency that view human nature as constituted by the social realm. Finally, we will look at the educational implications of economics as a discipline teaching a narrow, self-interested paradigm of agency. Rational choice theory stands to benefit immensely both in terms of prediction of future consumer actions and choices, and in explanation of economic phenomenon by recognizing the importance of influences on human agency such as normative commitments that exist beyond the realm of our mere desires.

Agency

Before beginning an exploration of rational choice theory and the philosophical tradition from which it comes, it's important to define agency more clearly, and why this philosophical concept is particularly relevant for economics as a social science. Understanding the intricacies of agency is an imperative first step before comparing the differing perspectives and assumptions made about its constitution. Economics claims to want to be able to predict choices made by agents, but this requires defining the process behind agents' decision-making. We must be explicit in defining an agent and their defining characteristics in order to evaluate the assumptions made by rational choice theory.

Agents, at the most basic level, are beings who act for reasons. This means that their actions can be explained by their mental states, such as their beliefs, desires, and preferences (even though they may not always be aware of their mental states). Agents are viewed as responsible for their actions because they are motivated to act by their own mental states. From this it can be said that beliefs and desires are thus *reasons for action*[3] (Reiss 2013, p. 13). This connection is integral to our definition of an agent because beings whose actions do not correspond with their mental states are almost entirely impossible to interpret. They are not intelligible as being responsible for their actions, because they are not being guided by any reasons.

Intelligibility is especially important when we look at agency from the third person perspective (as we regularly do in the social sciences), because others must be able to look at an

[3] I do not mean to confuse having reasons with being motivated to act. This does not mean that every time an agent has a reason to act, they do act, but it does mean that each action we observe is connected to an internal reason. There can be many competing motivations to act, but only one of them ultimately wins via the agent's behavior.

individual's actions and attribute reasons to them in order to understand them. When someone says that they understand why someone chose to do something, they are expressing that the action was intelligible as the agent's own action. If someone's actions were not connected to their mental state, and we could not attribute any reason for their action, then we would be unable to interpret them at all. In conducting research, economists seek to predict what consumers might do in a given situation and implicitly work to understand what reasons agents have for acting.

In following this line of thinking about intelligibility, Donald Davidson argues that any kind of rational cause for acting could be seen as an explanatory reason for an action (Davidson 2004). By explanatory reason, he means that we can say that this reason brought about the action. There is an importance here of a reason not just being linked to an action, but that it is rationally seen (by others) to cause the action. This understanding of explanatory reasons as rational causes is an underlying criterion for interpreting other beings as agents.

To further clarify, I think it's helpful to look at a hypothetical example.

If an agent, Joe, wanted a beer and remembered that he just recently bought a 6-pack of Budweiser, his action to go to the fridge and get a beer would be explainable by his desire for beer. If instead, when Joe went to the fridge, he poured himself a glass of orange juice, we would struggle to explain Joe's actions based upon our limited information about him. We would likely rather conclude that he must have changed his mind and decided that he wanted some orange juice instead. Joe's choice to get orange juice when we wanted beer can only be rationally explained by a cause such as 'he wanted orange juice'. Agents are defined by the connection between their mental states and the actions they take because of them. Rational explanation such as this is about the cause for the action.

As this situation with Joe above is explained, it appears silly and a little nonsensical to even entertain an idea where a human being gets up to satisfy a specific desire and instead chooses something entirely different for no reason[4]. This is the exact point raised by the discussion of intelligibility above. If someone is acting without connection to their own intentions, there is a barrier in understanding them because everything that we learn about their desires, hopes, or commitments are meaningless in regard to their action. We are intelligible to each other because this framework of explanatory causal reasons binds us together and allows us to make some baseline assumptions about other people as we interact with the world. Higher level expectations of others, such as holding others accountable for their wrongdoings or praising them for their accomplishments, are all built upon our understanding of each other as agents.

The point is that agency implies some minimal level of rationality in all agents. The word rationality is broadly used in the social sciences with different meanings, but I use it here to mean that someone is using reasonable or efficient means to achieve their desired ends. I will refer to this kind of rationality as **instrumental rationality**, as it is commonly referred to in other literature on the subject (Reiss 2013; O'Neill 2000). For a person to be considered an agent, we must be able to provide reasons for their actions, and these reasons must be in some way connected to the agent's goals or desires. In order for this causal connection to exist, we must be able to say that their action is tending towards some desired goal. Instrumental rationality in this sense is used as an 'instrument' to achieve their desired goal (by identifying the most effective steps to achieve the goal).

[4] Obviously, there are many situations where someone gets up to do something and then *changes their mind* and does something different. That is not the situation I am describing above. The situation in which someone switches their desires is completely intelligible to another agent, but they must rationalize behavior in order to fit in line with a cause. The point I am making above is that we cannot even make sense of a situation where an agent is doing something without a reason. We *fail to understand it*.

Instrumental rationality is a very easy assumption to make about an individual from the third person perspective because without attributing it to another agent, we cannot interpret them effectively. If I was just choosing to do an action with no goal or outcome or desire in mind, then how could a spectator understand any explanatory reason for my action? To further explore this subject, I first want to look at a schema that is outlined by Carl Hempel about rational agents[5] (Hempel 1965, p. 469):

> *A* was in a situation of type C
> *A* was a rational agent
> In a situation of type C, any rational agent will do x.
> Therefore, *A* did x

When we think of Joe, we would say that if Joe was a rational agent, and wanted to drink a beer, then he would go to his fridge and grab the beer he just bought this week. If we remove this 'instrumental rationality' from an agent like Joe, we may end up in a situation where Joe is pouring himself a cup of orange juice instead. He is in situation of type C (going to the fridge to get a beer because he desires beer), but instead of doing x (grabbing a beer), Joe does y (pours himself a glass of orange juice) The cohesive definition is that an individual exhibits instrumental rationality insofar as they take suitable steps to achieve their desired goals. This is precisely the type of rationality that Hempel is referring to in the above schema, and it is the baseline for what we understand as constituting an agent. In order for a being to be understood as

[5] Hempel states that the above schema captures how a "rational agent" will act, and I quite agree with him. However, I think that the term rational agency can be seen as quite redundant when the only expectation is that an agent takes efficient and effective steps to pursue their goals. Instrumental rationality is a part of understanding agency in general. Joe is not intelligible as an agent when he chooses to pour his glass of orange juice because he is not following any kind of instrumental rationality. His goal/desire was to get a beer from the fridge, and he engaged in an action that was completely counter preferential. It seems that if A (regardless of whether or not you define him as rational) does not do x in situations of type C, there is some difficulty in calling him an agent at all. By this I mean that if the agent does not have another reason for not doing x, then it is difficult for us to justify his actions.

having control over their own actions via their mental states, they must possess some kind of instrumental rationality.

Obviously, human agents are extremely complex and are filled with lots of competing goals and desires and thus we often do find ourselves in a situation where another human is slightly unintelligible. We typically attribute it to a misunderstanding of our own (i.e., we don't fully understand their goals/desires). By rationalizing, we attribute a mental state to someone, which they may or may not have, to make their actions intelligible to ourselves. In this sense, attributing instrumental rationality is part and parcel with interpreting another agent, and this is why we must come up with ways to attribute rationality to an actor when we rationalize. However, rationality is not the only element that defines human agency, we also possess a powerful ability to reflect and evaluate our desires. Agents evaluate not just what they desire and the consequences of pursuing said desire, but also whether or not that desire is worthy of pursuit in and of itself (Taylor 1976).

Harry Frankfurt, a contemporary philosopher, argues that the key defining characteristic of an agent is their second order volitions and thus their reflective capacity (Frankfurt 1971). It is important to emphasize that the unique ability for human beings to not be guided merely by their strongest desires, but to be able to reflect upon what is *worthy* of desiring is an essential aspect of agency. Frankfurt argues that the essential component of a agent is the control over their will, or their manifested action. Agents possess second order volitions, which is the ability to reflect upon the desires they have and decide which they would like to make their will. The critical difference between an agent and a wanton being is found in their ability to decide what they value and are committed to, and then change their will accordingly. Frankfurt suffers some criticism from his ultimate claim that there are just certain things that we *care about* and that is

what influences our second order volitions (Frankfurt 1982); but what I want to take away from Frankfurt here is that he points out the critical role that reflection plays in differentiating agents from other beings.

Reflection is what helps agents to decide which actions are worthy of taking, whilst instrumental rationality helps us to determine which actions are the most effective means to fulfilling those desires. From the third person perspective, people understand others through the relationship between an agent's actions and their mental states insofar as their actions are effective means to their desired ends. We are able to criticize, evaluate and commend people's actions because they chose them themselves. It is not merely a fulfillment of a desire, but a genuine choice made by the agent to make it their will. Taylor calls agents "strong evaluators" who are able to utilize a qualitative evaluation that is beyond merely 'is A better than B?' (Taylor 1976 p. 287). Rather, agents ask the question - is A worthy of pursuing as a desire in the first place? Agents are therefore much more complex than just purely weighing two desires against each other, but also in evaluating their desires and restricting their action if they think the desire is not worthy of pursuit (despite their desire for it). A good example of this would be choosing to be vegan, in which some individuals desire to eat animal products but choose not to because they believe it is better for the environment and therefore the right thing to do.

Understanding these intricacies of agency is an integral part to all aspects of the social sciences because the research conducted ultimately seeks to understand, explain, and predict human agents' actions and decision-making. By laying out the most basic picture of agency we can explore the different views that philosophers have taken to explain the constitution of human nature. The way that we each choose to understand agency has tangible impacts on the ways that we choose to interact with others and build our societies. Rational choice theory in economics

focuses on the major role that people's desires, or preferences, play on people's decision making and are not very interested in the way that people's preferences come about. This perspective of agency, being determined primarily by one's desires and the instrumental rationality that guides their pursuit, was powerfully developed by the philosopher Thomas Hobbes in the 17th Century. Philosophers such as Hobbes have been making claims about human action and the constitutive nature of human agency for hundreds of years. Looking at Hobbes' paradigm of agency helps to reveal how Rational Choice Theory follows a similar view.

Hobbes & The Rational Paradigm of Agency

Thomas Hobbes championed his views of rational agency in his work of political philosophy, *The Leviathan*. He viewed human nature as "innately selfish, competitive and distrustful" (Shafer-Landau 2020 p. 80) Hobbes argued that all agents have desires and aversions, and these are what ultimately motivate them to act. They choose to engage in actions that satisfy or help to pursue their desires and avoid the actions that would lead towards their aversions. For example, if I was scared of (averse towards) heights, I would avoid going into tall buildings and looking out the window. On the other hand, if I desired ice cream, then I would go to an ice cream parlor and get some ice cream. Hobbes' model of human action assumes that all agents possess instrumental rationality because they make decisions based upon which action would help them most efficiently achieve their desires. Hobbes emphasizes that there is no such thing as objective good and evil, or right and wrong; there are simply those things that we desire and things toward which we are averse. We subjectively define those things that we desire as 'good' and those things we are averse towards as 'bad'. Ultimately, rationality for Hobbes is regularly acting to maximize the satisfaction of one's goals and desires. Actions that go against one's own desires are irrational and go against human nature.

It is important to note that when we talk about rationality in terms of an agency here, the claim is that rationality is not about evaluating which desires, or goals, or pursuits are more valuable than others. We simply have them or do not have them. The only evaluation that rationality does is how efficiently one is achieving one's overall desires, rather than talking about the value of the desires themselves. However, this does not mean that rationality cannot provide *any* evaluation of our desires. Think about someone's decision to quit smoking cigarette - Hobbes would explain that their rational decision was made because smoking was going to

ultimately cut their life short, which would have limited the number of total desires they could satiate in their lifetime. In this way, reason allows us to critically evaluate desires in terms of the impact they may have on our overall efficiency of achieving our desires. All of these considerations do not tell us which desires we should value over others though. Some people continue to choose to smoke cigarettes despite the knowledge that it will cut their life short. Rationality is thus limited to only evaluating desires in relation to other desires, but not in terms of any qualitative aspect. As may be evident, Hobbes paradigm of agency is not interested in involving the evaluative capacities which Frankfurt and Taylor view as a central component.

David Hume was a classic thinker who agreed that our reason cannot help us to discern which pursuits are worthy or valuable (Hume 1748). He writes that reason cannot provide any knowledge about the quality of ends, or why some desires are intrinsically better than others. Rationality serves its strong purpose in helping us to discern which choices would be best to take, given the ends we seek to pursue. However, he believed that it is ultimately our passions, the way we feel about things, that motivates us to choose certain desires over others. He wrote that "reason is and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (Human Understanding III.iii.iii.2). An agent is thus seen by Hume to be guided by his instrumental rationality in how to efficiently pursue his desires, but these desires are shaped by an agents' passions. However, he also argued that agents were guided my altruistic and other-oriented desires, not merely selfish ones as Hobbes believed.

Hobbes did not see human agents possessing any capacity for evaluation and details a different account of human nature in his 1651 book, *Leviathan.* He describes the hypothetical situation of the state of nature, before any society has been created and there are no laws or rules of society. These humans exist in a world without any civilization and thus there is no

mechanism to enforce cooperation agreements - everyone is out on their own. Since there are no shared goals or objectives for human beings other than survival (and he views human nature and distrustful and selfish), Hobbes believes that human beings in this situation would only be motivated by personal gain. However, they may still want to cooperate because it would be in their best interest. Through cooperation, they could end up better off than they are now and satiate more desires by potentially trading with others for what they wanted more or by achieving goals that required more than just one person. Since there is no society, and thus no enforcement mechanisms for any kind of agreements, Hobbes claims that there would ultimately be no cooperation. None of the humans could trust each other and would instead exist in a constant state of war. By this he means that humans would always be at risk of being killed and taken advantage of by others. In his own words, the life of man would be “solitary, poor, nasty, brutish, and short”. Hobbes argued that human nature is so self-interested that, at the first opportunity, they would betray another person in order to maximize their own benefit. Since one cannot rationally trust the other, the only option you have is to ensure your own protection by acting first and trying to betray the other person before they betray you.

Hobbes lays out the problem that is faced by two agents who have come into contact with one another during the time of the first men. We can flesh it out for better understanding. Let's say that the first agent, Bill, has a surplus of berries and so would be willing to give them away in exchange for something that he doesn't have. The second agent, Jane, has an extra waterskin that she has made to store water, but is running low on food. It would make both Bill and Jane better off for them to trade, berries for waterskin. However, Bill would get even more benefit if he just betrayed Jane, took both her waterskins, and kept the berries for himself. The same could be said for Jane betraying Bill and running off with all his berries. The tough decision that is

faced by both agents is that the best payoff for each individual is for them to take advantage of the other actor and run off with their stuff. However, since both need the other person's item and have a surplus of their own, the best net outcome would be to just cooperate and completing the trade. For Hobbes, this creates a dilemma because neither Bill nor Jane will be willing to put their guard down to initiate a trade due to the risk of being betrayed, and thus no cooperation will ever occur.

Hobbes wrote of this situation prior to most of modern game theory, but what he implied (although using different language) is that the statue of nature is a prisoner's dilemma. The Nash Equilibrium solution to the situation, which is a set of strategies in a strategic situation for all players in which none of them would choose to deviate from their current strategy given that all other players keep the current strategy they are playing, is for both players to defect from their agreement and betray the other person. Hobbes did not use any of this terminology, but grasped the underlying motivations that are assumed in modern game theory. What Hobbes argues for is that rational agents would be motivated by their self-interest to betray the other person, but that the ideal outcome for everyone would be for cooperation. He understood that "by seeking to maximize self-interest, everyone is going to be worse off, in such dire circumstances, everyone is competing to gain as much as he can, at the expense of others" (Shafer-Landau 2020 p. 209). In the situation of Bill and Jane, neither of them would choose to deviate from betraying the other because if they chose to cooperate and let their guard down, their stuff would be taken, and they would be left with nothing. We will come back to the concept of Nash equilibria and game theory to show how this viewpoint of human agency persists through rational choice theory today.

Hobbes is emphasizing that the nature of human beings to maximize one's own desires is the biggest impediment to cooperation. His political theory culminates in the argument that for any society to be able to exist in the time of the first men, there must be two main functions in a society – rules that encourage or require cooperation and punish those who betray, and an absolute sovereign who can enforce and maintain these agreements. If a sovereign is in place, then agents will be able to cooperate without the risk of being taken advantage of by the other person. Absolute sovereigns cannot be influenced by the desires of taking advantage of others and would always act in ways to maintain fairness in interaction. This political theory is known as the social contract theory and began a long offspring of philosophical debate. The important aspect to glean from Hobbes' theories about humans and the story of the first men above, is that he believes all agency is ultimately guided from some notion of self-interest.

Hobbes' account provides what is called a folk psychological theory of human action. Folk psychology is the view that humans can and ought to be explained by looking at their beliefs and desires as evidence for their action. These explanations often struggle with prediction because of how complex the human mind is, and how difficult it is to observe someone's mental states. However, these types of explanations continue to provide strong explanatory power today for human agency. In the second half of the 1700s, Adam Smith began to build upon the previous ideas of political theory about individual action and how societies were compromised of individual decisions. Smith is widely considered to be the 'father of modern economics' because he took his theories of agency, rationality, morality, and politics and applied them to the economic realm.

In WN, modern economists interpret Smith as arguing that, as long as the society is just and organized, there is no need for planning of economies because everyone acting in their own

interest will culminate in good outcomes for society. He explains that when we go to the butcher for our dinner, we do not expect benevolence from them, but rather we expect them to do their job well because it is in their own self-interest (they will get paid for their work). We will look more closely at this passage later after exploring more of the contemporary debate about decision making. Smith is ultimately most famous in economics for this passage in *Wealth of Nations* that describes how everyone acting in their own self-interest leads to the best outcome for all, as if guided by an invisible hand:

> *He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By preferring the support of domestic to that of foreign industry, he intends only his own security; and by directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it. I have never known much good done by those who affected to trade for the public good.* (Smith 1776)

Smith's theory of human agency was also a folk psychological explanation and it related individual behaviors to the behaviors of society. However, economists needed more than this explanation for how economies function in order to be able to effectively predict outcomes and explain more complex processes such as where prices come from. Although Smith took large strides towards an analytical economic account, he left much to be desired in terms of precision and prediction for economists. Ultimately, many economists read Smith as being committed to a Hobbesian picture of agency (driven by our desires and self-interest), but cannot account for his prior work (TMS). Within TMS, Smith argues that human beings are socially constituted and genuinely care about how others judge their actions. By this he means that individuals change and adjust their behavior based upon the approval or disapproval that others in society give. Our desires, beliefs, and preferences are all shaped by our surrounding society and the reactions that

others have to our behavior. Rather than only caring about self-interest, Smith argues that we should appeal to *others'* self-interest in order to truly recognize and respect their agency. This leads to a very puzzling Adam Smith problem where economists view Smith as providing evidence for the power of self-interest in a market system and philosophers viewing him as a proponent of moral sentimentalism and advocating for the deeply social nature of human agents. This 'problem' will be resolved in a later chapter, but for now we will continue with the way economics portrays Smith.

As the social sciences expanded into the 20th century, scholars began to look to follow the path of the natural sciences by focusing on prediction capabilities and empirical tests as the main explanatory resources. In response to this push by the natural sciences, multiple disciplines began to attempt to mathematize folk psychological explanations to provide theories that were able to both explain and predict human behavior. Chief among these theories that arose from the Hobbesian picture of agency was that of rational choice theory in economics. The rational choice theory approach (which involves utility calculations and strategic evaluations) to economics is still very young in comparison to other fields, with most accounts being traced back to applications done by economists at the University of Chicago in the early 1940s (Herfeld 2018). The theory sought to create and apply mathematical formulas to support and ultimately predict the ways in which Hobbes (or Smith) saw human beings were motivated to act. These mathematical formulas were much better at prediction than folk psychological explanations were and could be used much more effectively as tools of study for a complex economy. As comes along with any effort to develop a model, rational choice theorists made assumptions about agency and narrowed down the scope to quantifiable properties. The model seeks to be predictively accurate despite failing to accurately describe the internal mechanisms of an agent.

Rational Choice Theory

Rational choice theory contends that agents are rational beings who seek to pursue their own preferences and act as if always maximizing their own utility payoff. This theory places economics firmly within the Hobbesian view of agency that we just described because it ultimately regards agents as driven merely by desires and aversions – captured through their 'preferences'. Rational choice theory is the predominant theory of decision-making in economics and hence is intimately connected to economics' conception of agency. The theory is commonly understood to provide explanation for an agent's action and predict their future decisions[6]. Since the usage of this theory is ultimately about individuals' actions, I consider this to be the mainstream view that economics portrays about human agency. This could be disputable as one could argue that rational choice theory is merely a positive tool to create economic predictions (Friedman 1966), but I contend that as a theory about human decision-making, it addresses

[6] Some rational choice and game theorists, such as Ken Binmore contend that the theory is a tautology rather than an argument (Binmore 1984), which means that it is true in virtue of the definition of its logical terms. This claim is certainly significant, as it means to compare the theory to mathematical theories and contend that rational choice cannot actually be proven false because it does not say anything substantial. Binmore and other rational choice theorists who hold this view believe that the claims are merely true in the sense of the meaning of the theory. There are larger concerns about viewing rational choice theory in this way but on a surface level, this type of argument doesn't do much good for rational choice theory. *Prima facie* this doesn't even appear to be true, as the claims of rational choice theory are not descriptive as in a tautology, but rather make an argument about how they believe agents make decisions). In making predictions, rational choice theorists put together a set of propositions that formulate an argument such as:

Rational Agents always choose the option with the highest payoff
Option B has the highest payoff
Therefore, Rational agents will always choose option B.

This argument can clearly be proven false if we could determine that, despite the fact that option B is the highest payoff option, the rational agent selected option A. This is not a tautology, as the statement "x=y or x≠y", but rather a theorem that can be tested and proven false (Guala 2006). I acknowledge that these views of the theorem exist, but will define, describe, and evaluate rational choice theory in view of it as an argument which can be true or false. I believe that the importance of the theory comes from the fact that it is a significant argument about agency, rather than being a descriptor of the world.

questions about action, and thus has something to say about how we define an agent. A tool to predict decision making has assumptions built into it that contend for a certain interpretation of agency and therefore even if it acts as a tool, it also works as a paradigm of agency. Although we will be looking at how it is used in economics to look at decisions made by consumers in a market setting, it should be noted that the theory is widely applied to any social study that looks at actions and choices of an individual. The theory has been used in political science, international studies, sociology, anthropology, and many other disciplines that examine human behavior. It emphasizes verifiable behavior of an agent in order to make prediction and explanation within the social sciences more 'scientific' (in relation to the physical sciences) via mathematizing the folk psychological understanding of agency. Ultimately, the purpose of applying mathematical formulas to the claims about human agency that were made by Hobbes, Smith and other classical thinkers is to be able to provide more precise predictions and causal explanations about action that don't rely only on unverifiable mental states.

Rational choice theorists argue that human beings are innately rational and make choices by maximizing their own payoffs. Therefore, they always select the alternatives that give them the best outcome for themselves, which is determined by their personal preferences. Instrumental rationality is the key element that the theory is built upon, as each individual's actions can be drawn back to their personal goals and explained in light of how that action helps to achieve their given goals. Any action that an agent takes undoubtedly helps them to further pursue one or more of their desires, because, under this theory, that is how agents operate. Rational choice theory is taught and applied in economics as a global theory of rational decision making but, since individuals' mind states are complex and difficult to regularly interpret, economists are only interested in using empirically verifiable data for their models. This means we must rely

upon people's observed behavior to give insight to one's underlying beliefs and desires. An agent's choices serve as a representation for their preferences for different items because instrumental rationality guides them to choose the action which would bring about their desired goals.

Neoclassical economists sought to push economics towards focusing only on observable behavior to explain agents' choices rather than looking to their mental states (which could not be verified or seen). They believed that they needed to push their respective disciplines much closer to the way the physical sciences looked at the world and away from folk psychology. Looking to beliefs and desires, as folk psychology does, to explain an agent's action does not provide verifiable states as explanation. This type of explanatory power does not require many inputs, beyond an agent's choices, and yet still can help us to predict what they might do in the future.

Expected Utility Theory

The formal definition (with more economic jargon) of rational choice theory is that human beings are rational agents who act as if they make decisions based upon their preferences in order to act as if they maximize their expected utility. Obviously, some terminology warrants explanation here. An agent's preferences are a ranking of their alternatives based upon which options they prefer more. They are implicitly made up of an agent's beliefs, desires, and tastes about the world around them, but they are captured from their observed actions. For example, I prefer iced tea to hot tea. Given the choice between iced tea and hot tea, I choose an iced tea, revealing that I prefer one to the other. These preferences are obviously contingent upon my environment, mood, social structure and a variety of other factors, but all of these factors are exhibited in my action to choose one alternative over another. I could be more specific and say that I prefer iced tea to hot tea all things considered, but if it's a very cold day out, then I prefer

hot tea. This framework of preferences influencing our choices allows for the inclusion of a wide variety of factors to be exhibited via our choices.

John Von Neumann and Oskar Morgenstern are credited with the foundational work on utility theory in 1944. Their theory begins with taking "preference" to be synonymous with choice because an agent's preferences are revealed via their choices (Von Neumann & Morgenstern 1944). The choice that an agent makes takes all factors into account, and tells us that *ceteris paribus*, they prefer the chosen alternative the most. Rational choice theory proponents, such as Milton Friedman, argue that using preferences in this way brings more simplicity to the model and avoids having to deal with unverifiable mental states without losing out on any significant accuracy in prediction (Friedman 1966). This is the key explanatory benefit that comes from focusing on behavior to explain an agent's action and predict future decision making. You need less information to reveal preferences under this model of rational choice.

Von Neumann and Morgenstern utilized this theory of revealed preference to build a mathematical model which could measure what the expected payoff of an agent's alternatives were. They came up with conception of utility which essentially explained that when individuals were faced with risky prospects (meaning that there was a probability associated with each outcome), they will act as if maximizing their expected outcome over the specified course of time. Utility is the total personal benefit, or payoff, that one receives from making a given choice. Human agents would like to be better off rather than worse off, and thus will make decisions that are maximizing to the completion of their own desires. Utility is often measured and easily understood in numbers, or even in dollars. Think about someone given the choice between a $10 bill or a $1 bill, no strings attached. Anyone and everyone would choose the $10

bill because 10 is more than 1. When looking at more complex situations, you could imagine making the choice between a 50% chance at $100 and a 90% chance at $10. Expected utility theory explains that rational agents will multiply the probability by the payoff to solve for the expected utility. In this case, any individual should choose a 50% chance at $100.

Outside of monetary concerns, you can think of utility as an index of all of your preferences. We prefer to acquire certain things over others, according to the desires that we have. Utility is a relative measurement of benefit that an agent receives by choosing an item or action that he prefers more than another. This does not mean that it is necessarily a differential benefit, but rather that the total benefit is determined in relation to other choices. For example, there are some items that are substitutes for other items, such as margarine or butter. An individual may prefer one item to the other or may be indifferent between them, allowing them to be substituted for each other easily as they both return the same amount of utility to the consumer in this case. Von Neumann and Morgenstern claimed that one could use an expected utility model to predict the choice of agent even for more complex decisions such as major life changes or even moral choices. They proved all of this through mathematical theorems alone.

Assumptions of Rational Choice Theory

In order to build effective predictive models, economists generally assume an image of agency that is ideally rational in the sense that it is how a human being would act if there were no deficiencies or distractions to their rationality. There are four core axioms of an economic agent that are commonly shared and described similarly in economic textbooks' description of rational

choice[7]. Although there is some variety in the literature, these four axioms capture the essence of the view of agency that rational choice theory assumes.

The first axiom is that an agent must have a known set of alternatives, which means that they know what they can and cannot do in the situation. This axiom seems simple, but it is important when we want to talk about available information to agents. Within most models of consumer choice, perfect information about their choices is assumed for the actors in order to simplify the model and reduce the amount of knowledge needed for consumers to reasonably participate in the market. This information is not necessarily all understood by the consumers, but the assumption comes from the belief that most consumers know basic information such as how much a carton of milk, or a dozen eggs should cost in their market, and thus can detect price changes. Rational choice theory holds that agents are aware of their most likely available choices when faced with decisions which allows them to reasonably rank alternatives in comparison to each other. If you weren't aware of the options that you could choose from, it would be impossible to discern preferences between unknown alternatives.

The second axiom is that an agent has preferences regarding the relationship between any two given alternatives. Given the choice between A or B, they will either prefer A over B, B over A or be indifferent between the two. This means that whenever someone is faced with a situation where they can make a choice, they have an opinion on the relationship between all of their possible choices. If this weren't the case, then people would not be able to reasonably decide between their alternatives, because they would have no way of ranking them. Without a preferential ranking, agents could not recognize the alternative that gives them the highest payoff

[7] You can look to many different sources on what the main assumptions of rational choice theory are, but here are few from defenders and critics alike; (Green 2002); (Levin & Milgrom 2004); (Simon 1990); (Reiss 2013); (Von Neumann and Morgenstern 1944); (Kahneman and Tversky 1979)

and thus could not exercise their rationality. When someone is indifferent between two choices, it means that they need just a small nudge to pick one choice or the other. If Joe was indifferent between oranges and grapefruit, he could be incentivized to buy one or the other purely by one of them being cheaper than the other (if he cared about saving money). This does not mean indifferent in the sense that he has no opinion about which one he chooses, it means that he prefers them equally, all things considered. Indifference is still considered to be a preference and therefore allows the alternatives he is indifferent towards to be considered in relation to other alternatives as well.

Third, an agent's preferences are transitive; so, if someone prefers A over B and B over C, they must necessarily prefer A over C. Transitivity simply means that we apply logic to our preferences when comparing different alternatives. It makes sense that, if I ranked chocolate over vanilla and vanilla over strawberry, given a choice of ice cream I would choose chocolate. Whether the choice was reduced to chocolate and vanilla, or chocolate and strawberry, chocolate would always be chosen because I prefer chocolate the most. Transitivity is one of the most important elements of a rational agent in this context, because it allows for comparison among alternatives. Without transitivity, preferences could not be truly "preferences", as it would destroy the essential element of preferring one alternative to another. Transitivity allows for utility to represent a hierarchy of payoffs and preferences, and hence gives us the pathway to selecting our most preferred alternative.

Finally, a rational agent will always select the most preferred alternative from their set of options in order to maximize their overall payoffs. If given the choice between A and B, and they prefer A to B, they will then choose A because that is what they prefer the most from their given set of alternatives. This is the final culmination of the view of agency that comes from

mainstream neoclassical economics. Agents are viewed as maximizers because that is how a rational agent operates according to rational choice theory. Given the three axioms above, the only **rational** choice when faced with a given set of alternatives, is to choose the one that gives the highest payoff. This view aligns clearly with the view of agency discussed by Hobbes - humans are guided by their desires and thus decide how to act in their own best self-interest. Essentially, rationality in rational choice theory is *always selecting the option that is most preferred from a known set of alternatives*, and all agents are rational beings who are guided by this principle.

So, the four pillars of the economic rational agent are information, preferences, transitivity, and maximization. Let's put a very easy hypothetical example to this to see how all four pillars can be applied to a given situation. Let's say that Sarah went to the grocery store and wanted to buy some produce, so she is making a choice within the produce section of the store and knows what her options are (or would be able to easily collect this information by walking around the section). We assume that Sarah has some kind of preferences between her available choices of produce, and, if asked, could tell us which vegetables or fruit she prefers to others. These preferences lead to her having a 'best' option, and we assume that she wants to choose the item from the produce section that she most prefers because she is a rational agent. Sarah is going to maximize her own benefit by choosing to buy apples because apples are her favorite item in produce. Maximization means that Sarah is always going to decide, given her information and choices, to buy a produce item that she likes more than all the other options. She will not select oranges instead of apples because she prefers apples much more. Economists do not claim to be able to tell you what option is the "best", but they can tell you how a consumer will act given their environment and constraints.

As we said earlier, preferences are hierarchal, and alternatives must be able to be compared to each other via their respective utility. Rational choice theory argues that this type of utility index can be inferred from agents based upon their past behavior because agents will always choose the option that brings them the highest payoff. Obviously, agents may be subject to certain constraints, such as the amount of money in their wallet, but they will still make decisions based upon their preferential ranking. So, rational choice theory assumes that all of our differing beliefs and experiences are captured through our observed behavior and the choices we make, which helps to inform ourselves and others of our preferences. In reducing the complex human being down to this "economic man", to borrow the term from Herbert Simon (Simon 1955), economists are able to utilize rational choice theory to predict the decisions that consumers will make in an economic setting.

The reason this assumption of rationality is helpful in economics is because it imposes constraints on the possibility of choices that consumers are making. They assume that *on average* consumers will follow this pattern of decision making and therefore one can make predictions for their future actions once we have collected data of their previously observed behavior. Economists are not interested in shining light on the black box of the human mind if it does not help to predict their actions. They would rather look at the way that agents actually act (observed via their behavior). As I said earlier, an agent is defined by their action and reasons for that action. Their actions are connected to their mental states, such as beliefs and desires, and this connection is what makes an agent able to be observed and interpreted by others. An agent is typically understood to have some level of instrumental rationality, which is the concept that if an agent desires *x* and action *a* is means to achieving or acquiring *x*, then the agent will do *a*.

This global understanding of rationality is fundamental to the way that rational choice theorists see human agents.

Its evident that this view of agency being based on reason comes from the work by Hobbes who argued that human agents are guided only by their desires and aversions. Furthermore, rationality is not meant to tell us anything about what we *should* desire, only how we should go about *achieving* those desires. Hobbes believed that human beings were inherently rational and self-interested, and thus they were motivated to act in pursuit of their desires. Rational choice theory similarly argues that agents, based upon the preferences that they have, make decisions that lead them to achieve or acquire the things that they prefer most. Preferences are not evaluated by rationality in any sense though. They are considered to just be constitutive of the agent, and the theory does not intend to address where the preferences come from. Expected utility theory is useful when looking to describe individual agents' choices on their own but needs more to account for payoffs in a social context. Human beings are always operating with their society, and therefore rational choice theorists must utilize additional tools to address this added variable.

In the real world we must regularly make choices that are not certain and *depend upon other people's choices*. You can think about some of the simplest examples of this such as when you are at a stop sign and it is your turn to go. Maybe there is another car at the stop sign to the left of you, and even though it is your turn to go there is a chance that they might also go. You make a decision that incorporates the chance that the other driver chooses to go and potentially hits you. Your outcome depends not only what you choose to do, but also what he chooses to do, and therefore there are different probabilities associated with multiple outcomes (it's unlikely that he will break the social norm that says that you will go first since you arrived at the stop sign

first, but it is possible). Probabilities help theorists to explain the fact that agents do not have complete control over the outcome of their choices, only the decisions they make.

Von Neumann and Morgenstern not only developed the concept of expected utility in general, but also expanded it to strategic situations. This work by Von Neumann and Morgenstern laid the groundwork for the development of what is now known as Game theory. It is important to recognize that the field was started by a mathematician, with the emphasis being placed on precision rather than explanatory value. Looking at game theory through this lens reveals the benefit of making what may initially seem like restrictive assumptions. The study of games helps to build out rational choice theory to include agents acting in specific social contexts in which their outcomes also depend upon other agents' choices. Game theory gives economists and other social scientists a further set of methodological tools to look at agents within a strategic context. It can be understood as the application of expected utility theory to situations that involve more than one agent.

Game Theory

Utilizing the 'economic man' that is formed in rational choice theory, game theorists argue that players in strategic social situations follow the same rational decision-making and guidelines. Rather than just having to be concerned about their own choices, actors must also be cognizant of other agents' choices. This is not only useful for prediction purposes when we look at strategic scenarios, but it can also be seen as a further elaboration on the theory of agency within economics. Primarily, it details how an agent's choices could change in the context of other agents. Since agents do not operate merely on their own, as an individual, anyone seeking to understand agency must simultaneously recognize the influence that others and society has on the individual. Ironically, this criticism is exactly what Hobbes received from the likes of Smith about overlooking the influence of society on the individual. However, as we will see, the fundamental assumptions in game theory still view agency from the Hobbesian picture with cooperation being a byproduct rather than a genuine intention.

Game theory is used as a modeling tool to show the best possible outcome when agents are faced with a strategic situation. These situations are strategic because they have strategic interdependence, which means that the best alternative for an agent depends on what another agent does (Harrington 2015). Game theory focuses more on the effects of other players choices on one's own decision making, something that expected utility theory only does implicitly through preferences. You could argue that one's preferences (and final choice) involve the influence of other people's action, but game theory studies the changing effects of one's payoffs based upon what strategy (or alternative) an agent chooses to deploy.

As I discussed earlier, Thomas Hobbes held some of the same views of rational agency that are displayed in game theory. The concepts outlined in the Hobbesian state of nature has evolved through the more formal work of game theorists to now be understood as a simple example of a prisoner's dilemma. The self-interested nature of agents leads them to a Nash Equilibrium solution that is bad for both of them. Remember, Nash Equilibrium just means that neither player would be willing to change their strategy given the strategy of the other players, and thus they are at an equilibrium. Although this may not be the most optimal outcome, the players have no incentive to switch their strategy because they would be worse off if they did Although there is no explicit reliance on selfishness like there is in Hobbes, Game theorists follow the Hobbesian conception of human agency through their expectations and assumptions of rationality.

Within traditional game theory, all agents are assumed to be rational, which means that they act in their own best interest by selecting a strategy that, based upon their belief about how other agents will behave, will maximize *their payoffs*. Furthermore, each player assumes that the other player is also rational and thus will be maximizing their own payoffs. This view of agency comes from the same foundation of rational choice theory with agents as fundamentally calculative, methodical, and self-interested beings. It is assumed that agents faced with a strategic situation are often (but not always) able to recognize their possible alternatives and the respective expected payoffs that come along with them. This makes no claims about the type of interests that people have, allowing a wide range of interests to determine respective payoffs just as we discussed with preferences in rational choice theory. People's payoffs are determined by

their preferences. Rationality in game theory just means that people will pursue whatever is in their best interest, no matter what those interests may be[8].

Hobbes' hypothetical situation modeled the central ideas of what is now known as the prisoner's dilemma. The situation that is created in game theory is that of two criminals who commit a crime together and are caught by the police. The police separate them and individually ask them to confess about the other's crimes in exchange for a shorter prison sentence for them and a longer prison sentence for their accomplice. The police don't have enough evidence to convict them for the major crime, but even without a confession, they could put them away for a 1 year. The dilemma is that neither criminal knows what the other person will do, and they could be better off betraying their partner and talking to the police to get a shorter sentence. They would be best off if they talk while the other person stays silent, as they would get no time in prison. However, if they both confess, then the police will have enough information to put them both away for 8 years. This can be illustrated through a simple table:

[8] I want to clarify the nature of the assumptions baked into rational choice theory, and game theory, which is that it always comes back to the *agent's utility function*. You can only understand altruism in this model as a benefit to the individual who engages in the altruistic act. There is no such thing as 'pure altruism' in which the agent increases the other player's payoff and gets nothing in return. It's not conceivable because an agent must always select the option with the highest utility, and thus an altruistic action must be increasing one's utility in order to make them select that option. This makes the model committed to the Hobbesian picture because it is always related to self-interest. That doesn't mean it is a selfish model, rather it is self-interested.

Xavier	James		
		Talk	**Stay Silent**
	Talk	8 yrs, 8 yrs	0 yrs, 10 yrs
	Stay Silent	10 yrs, 0 yrs	1 yr, 1 yr

The numbers represent the number of years that each criminal would have to serve in prison, with the orange numbers representing Xavier's payoff and the blue as James'. The columns represent the strategy employed by James; The rows represent the strategy employed by Xavier. As you can see, cooperating is a better payoff than both choosing to betray. However, they will both be incentivized to betray because they want to get the fewest years possible, hence the green shading representing the Nash Equilibrium solution.

Something that Hobbes was keying into in in his hypothetical situation of the state of nature is that the best possible outcome for society as a whole was cooperation because it had the best possible net payoff. As you can see in the prisoner's dilemma modeled above, the same thing is true - the lowest total number of years served in prison is actually achieved by both of criminals staying silent. However, both agents individually receive 1 year and thus it is not their best individual outcome - they could choose to talk when the other cooperates and receive 0 years. The Nash equilibrium for this game is both players choosing to talk. The reason for this is because the dominant strategy in the game is to talk. This means that talking is the best strategy no matter what the other person chooses to do. If Xavier chooses to talk to the police, James should also choose to talk because 8 years is less than 10 years. If Xavier chooses to stay silent, James should still talk to the police because 0 years is less than 1 year. Hobbes recognized that this was the paradigmatic example of human interaction in the state of nature, and that's why we

need a social contract to enforce cooperation. He did not view cooperation as possible without rules and a sovereign to enforce them.

Game theory has obviously evolved a lot since Hobbes and as a discipline through the 21st century worked include a huge range of studied games such as imperfect information, private information, signaling games, cheap talk games, repeated interactions, etc. All of these different variations of games try to examine a particular unique aspect of human decision making (such as when an agent is faced with an unequal information distribution and does not know what someone else desires). As more experiments have been conducted and games have been examined based upon the axioms of rational choice, game theory has evolved to be able to suggest how outcomes will change as an agent's strategic environment changes. This pushes the boundaries of rational choice theory as it theoretically allows it to be applied to a wider range of situations when explaining or predicting an agent's decisions.

For example, the prisoner's dilemma traditionally assumes that the highest individual payoff is given to the agent who chooses to betray when the other player chooses to cooperate, which is why there is a suboptimal Nash equilibrium. This is because the preferences of the agents are set up so that they care most about their own personal outcome and benefit and have no concern for a benefit to someone else. However, in some cases, the highest payoff for both parties may be to cooperate, as they get some added benefit from cooperating with the other person. Game theorists today recognize that some people's payoffs involve actually choosing to cooperate because it brings a better payoff for an agent who is inclined to help others. In this way, game theorists can actually choose to explain payoffs that may appear to incorporate other-oriented preferences as merely driven by a self-interested preference to feel good about the equality of a payoff (Woodward 2009). Game theory evolved to attempt to not only address the

challenge of altruistic behavior, but also explain its existence in studied games by changing the image of the underlying preferences of the players.

In added complexity, game theorists also try to account for the varied outcomes between single-shot and repeated games. Agents respond differently when they know that they will have to play a game again with a player, rather than if they know that it is just a single round. Testing within a single shot game versus a repeated game helps to reveal how much of a factor the number of times a game is played affects the strategies that players employ. The theoretical considerations of game theory have expanded beyond the idealized version of games that assume each player has perfect information about the other player and thus have opened the door to applying theory to more realistic and complex situations. It is important to recognize that there are some great examples of game theory being applied to the real world in order to better address engineering of an institution, system, or mechanism.

The Nobel Laureate Alvin Roth has worked to utilize market matching systems and game theory techniques to build organ matching designs that more effectively match kidney donors with patients who need them (Roth et al. 2004; Roth et al. 2006). Roth and other researchers use matching techniques from game theory to overcome the common problem within kidney pairing in which recipients are not compatible with their assigned donor. By creating pairs and 'chains' of donor-patient pairs, kidney exchanges are able to be more efficiently allocated by swapping one of the incompatible partners with a compatible partner in a different pair. In this way, 2 pairs of incompatible pairs swap partners and become 2 compatible kidney exchange pairs. This work, alongside some school choice design, earned Roth his Nobel Memorial Prize in Economic Sciences in 2012. The reason I bring up this example is because it is a unique intersection

between moral and efficiency considerations where game theory techniques were able to provide a significantly better outcome than before.

This example helps to avoid the common misconception that rational choice theorists only care about efficiency or are only focused on money. If you asked an economist what the most efficient method for allocating resources would be, they would tell you it is a market system. It would not be ridiculous to assume that the response that economists would give to the problem of organ donations would be to just let organs be sold on a free market. They would argue that this would be the most efficient way to allocate them, so that's the route we should take, morals be damned[9]. However, Alvin Roth showcases the importance of finding efficient solutions within the moral boundaries of the society that we live within. In this case, that means not allowing body parts to be sold on the open market but instead looking for different solutions. Roth's organ market design has saved countless lives and has made careful consideration of the moral constraint. Game theory helps us to be able to engineer some significant systems within our economy, utilizing the assumptions that are made by rational choice theory to evaluate agent's decision-making. There is a significant ongoing debate though between the role that theory in itself plays, versus the benefit from laboratory experiments that test the assumptions of game theory and rational choice theory with real agents.

[9] Look at Mohammad Akbarpour talk about this in an interview with Stanford Business here: A Beautiful Application: Using Economics to Make Kidney Exchanges More Efficient and Fair | Stanford Graduate School of Business

FCC Spectrum Auction

In order to better understand this debate and why it's important for economics' conception of agency, let's look at an example of game theory in action. One of the most famous examples of game theory's success in prediction and application to the real world is the FCC Spectrum Auctions in 1994. Auctions are a great example of strategic situations that clearly exist in the real world and are also studied within game theory. Auctions are strategic because each bidder's outcome depends not only on their own actions but also on the action of other bidders. Auctions are also really helpful for measuring people's preferences within game theory because you can strictly define preferences to a bidder's willingness to pay for an item. The value that they place on an item inherently includes all of their beliefs about the item and their desire for it. This helps to overcome a consistent struggle that rational choice theory faces – determining an agent's preferences.

Auctions can take many different forms, such as a first price auction, where the highest bidder wins and pays the price that she bid, or a second price auction, where the highest bidder wins and pays the price that the second higher bidder bid. Theory comes from studying the decisions that are made by individuals in auctions and is extrapolated to more general conclusions about their behavior based upon the rationality assumptions that are made about agency. For example, theory tells us that in second price auctions, the dominant strategy for players is to bid their true value because they would get value from winning by paying the second highest bidders' price; but in first price auctions, bidders will bid slightly under their true valuation in order to get some value out of winning since they would be paying the price they bid. These types of considerations are all important for building auctions in the real world, and that's why game theorists were solicited for advice when devising the FCC auctions.

In the 1990s, the U.S. government wanted to begin auctioning off all of the radio frequencies that they weren't using for public means to private institutions and individuals. The FCC was in charge of setting up an auction that would account for a wide range of regulations and specifications with a few key goals in mind – efficient use of the spectrum of frequencies; encouraging development of new technologies; avoiding excessive concentration of ownership; and accounting for minorities/underprivileged groups during distribution of licenses (Alexandrova & Northcott 2009). They asked for public comments in order to develop the best possible design, and this was one of the first revolutionary moments for real-life application of game theory. Economists and game theorists that were hired by the FCC, and some who were just interested in the project, sent in their comments, and participated in rounds of meetings and planning.

The final rules and regulations of the auction were covered in a document longer than 130 pages, which was distributed to everyone interested in bidding on the spectrum. The auction utilized rules that had never been seen before in a live auction and had only been proved through theory and laboratory testing. The auction took the form of a simultaneous multiple round auction which meant that licenses that had similarities were placed into bidding groups and all groups were auctioned at the same time. The individual licenses were bid on over multiple rounds, with the results of the previous round and the minimum bid to remain in the auction being revealed before the next round began. The auctions were a dramatic success and widely applauded as a fair and effective distribution for the radio frequency spectrum. Multiple countries prior to United States had tried to auction off their respective radio frequencies with limited success based upon their intended goals. However, The FCC raised $20 billion, and the licenses were distributed in an efficient manner that allowed for bidders to change strategies

based upon information revealed as the auctions evolved (Cramton 1998). My focus here is not on the specific rules that were detailed, but rather the methodology of coming up with the rules.

Some philosopher-economists have used it as an opportunity to even explore what it means to progress within a social science like economics (Alexandrova & Northcott 2009). For our interests here, I think it is a perfect example to explore where the strengths and limitations can be found in rational choice theory. The theories about human decision making that are found in rational choice theory were incorporated into the regulations that guided the FCC auction in the form of assuming that each bidder wants to maximize their expected utility and will act rationally in pursuing that goal. Economists in favor of the view that game theory played a central role in designing the auctions argued that the FCC "chose an innovative form of auction over the time-tested alternatives (like a sealed-bid auction), because theorists predicted it would induce more competitive bidding and a better match of licenses to firms" (McAfee & McMillan 1996 p. 160). Obviously, the contention of McMillan and McAfee is that using theory was the primary cause of the success of the auction as opposed to merely looking back at history.

This claim is important to the defense of theories like rational choice and to the economic discipline as a whole because it seems to imply that what is typically being done in policy making is looking back to the 'time-tested alternatives. For example, Ben Bernanke (who recently won the Economic Nobel Prize for his work), the chair of the federal reserve during the 2008 crisis, actually relied heavily on looking back to the tactics that were used by the government during the Great Depression as opposed to any new economic theory. His research on the Great Depression revealed that bank failures actually cause downturns rather than merely being a side effect, which led him to choose to push for bailing out many of the failing banks during the crisis (Bernanke 1983).This paints economics as a discipline that is primarily

informed and guided by history, rather than any real kind of theory about how an economy works and agents act. However, proponents of theory like McMillan and McAfee seem to believe that this is not the only avenue that is available to economists, and that we can genuinely provide explanatory and predictive analysis for agents' decision making by looking to theory. The success of the FCC auctions certainly supports their claims that theory has a real and powerful place within economics and their conceptions of agency.

It should be recognized though that as theorists were contributing their own beliefs about how different situations would affect the auction, there were simultaneously numerous experimental tests being run of auctions in order to see how different rules and restrictions affected people's decision-making process. There were practical concerns to address within the application of the theory that rational choice and game theory proposed. McMillan argues that these types of practical concerns ultimately came down to judgement calls that had to be made about how best to bring about the desired result given the theory we knew. However, as Alexandrova and Northcott point out, this response doesn't tell us anything about the reasons why specific judgement calls were made. Economists needed to make decisions about how theory should be applied in the real auction set-up, but theory could not inform these kinds of judgement calls (Alexandrova and Northcott 2009). Experimental tests gave the possibility to look for evidence that could influence judgement calls in one way or another.

In other words, rational choice theory was just one piece of the puzzle in setting up a successful FCC auction. Theory laid the foundation for how certain elements of the auction would work and highlighted the issues that needed to be addressed in the practical application. Yet, there was no overarching theory that could guide how an entire auction works, and thus work had to be done by the FCC and economic advisors to piece together the results of specific

experiments into a complete picture. The FCC Spectrum Auction was undoubtedly a wild success for theorists to argue that their theory was practically applicable beyond just an abstract observation. However, it also highlighted rational choice theory's heavy reliance on behavioral economics to bring the abstracted theory into line with reality. We can not only appreciate the power that theory granted to the economists working on building the auction system, but also recognize that there is a limit to theory's explanatory and predictive power.

The FCC Spectrum Auction is a powerful real-life example of how rational choice theory and behavioral economics worked together to be able to predict agents' decision making process and regulate them in a way that led to an efficient and desired outcome for the auction. However, despite the success that came from the auction model, they also highlighted that rational choice theory was only able to provide predictive power within the very specific constraints of an auction setting. Every finding from an experiment told researchers how individuals would act given all the constraints of the environment that they were placed, whereas theory required abstracting assumptions that made the environment less impactful. This is precisely the problem that rational choice theorists face. When theories are formed because of historical or experimental evidence, they are limited to the specific context that they were based upon. There is limited application that can happen from rational choice theory, because of the complex social nature of economies. In this way, rational choice theory's predictions are only accurate given the idealized context that they assume. Increasingly, the work done by psychologists and behavioral economists has cast doubt on the depth of this predictive and explanatory value that rational choice theory gives us in economics. Looking outside the discipline of economics to research done by psychologists reveals deeper inaccuracies in the assumptions made about agent's rationality.

Actual Human Behavior

Before I address the psychological inaccuracies of rational choice theory, I want to take the time to further address arguments about rational choice theory as a theory. Up until this point, we have been regarding rational choice theory as a paradigm of decision making and implicitly, agency. I argued that the theory has something to say about the psychology of an agent because it argues that there is a specific calculative process that will guides choices (appealing to the internal psychology of an agent). However, not all rational choice theorists would accept this classification of the theory as psychological, rather some argue that it is actually more similar to a theory which is focused merely on consequences rather than internal causes of action (Satz and Ferejohn 1994). This argument emphasizes that one's environment is the most important factor for decision-making and that rational choice theory is only focused on external, empirically verifiable states therefore psychological criticisms do not apply.

Prima facie, I think it is extremely plausible that our preferences and mental states are informed and constrained by our environment. However, that is not the same claim as arguing that therefore you do not have to be concerned with the psychological factors of an agent. The debate over the influence of environmental factors on our psychology, does not provide a strong groundwork to claim that we are ONLY a product of environmental factors. First, environmental constraints do not on their own cause action of an agent. If an individual goes to the grocery store with $20, that constraint does not tell me anything about how they will act and how much of that money they will choose to spend. The limitations tell me what they *are able to* buy, but no what they *will buy*.

Let's look at the common economic argument that companies seek to maximize profits in a competitive environment - there are two ways to view that type of statement in terms of what it tells us about agency. You could say that the people who run the companies are guided by their rational decision making to find the best way to generate profits and that it is thus intentional profit-seeking behavior. You could also say that the environment that the people exist within, a competitive market system, is what causes their company to maximize their profits because if the individual people weren't able to do that, then their company will go out of business. This second viewpoint makes no appeal to the psychological states of individuals who are in charge of the company, just explains their action via their environment. Now there's a lot of debate around whether or not firms actually work to maximize profits, but let's pretend for the time being that they do. It is not doubtful that the competitive environment has an influence on why they would be incentivized to maximize their profits, but that is a different claim from saying that companies maximize their profits only because of their environment.

Rational choice theorists are bound in some sense to recognize the psychological aspect of this claim because otherwise they are unable to extrapolate the casual reason for agent's choices in a competitive market system. They lose a lot of explanatory value about an agent's decision making, and even if rational choice theorists claim they're only interested in prediction rather than explanation, their prediction relies upon their ability to draw conclusions about the preferences at play (Herrnstein 1990). By this I mean that it is a significant statement to say that companies are maximizing their profits in a competitive environment because they can recognize the profit maximizing function and that it their highest given preference. However, if one appeals merely to their environment, we lack any kind of explanatory or predictive capabilities to extrapolate about companies outside of a competitive environment (which many companies do

not operate within). If there is any part of an agent's decision-making process that relies upon some appeal to their inner psychology, then rational choice is open to psychological critiques.

I think that rational choice theory must recognize and incorporate the significant psychological shortcomings in its theory in order to seriously participate in the agency and decision-making debate. These criticisms are not minor and at first glance prove very difficult to overcome without appealing to some kind of interpretation as Satz and Ferejohn outline. Rational choice theory can develop stronger predictive and explanatory capabilities if they integrate a more realistic psychological framework. Now that we've explored the theory in full and examined its broad applications and axioms, let's look to the challenges that I have been alluding to in terms of psychological accuracy. As I will show in future chapters, it is these psychological inaccuracies which lead the theory to fail to accurately predict decision making in the laboratory setting.

Prospect Theory

Daniel Kahneman and Amos Tversky are two psychologists who have been fundamental to the testing of rational choices theory's psychological axioms. Both researchers who study decision-making, their work and empirical findings formed the basis of explaining economic decisions in light of all that we know from psychology about how individuals really make choices. Kahneman received the Nobel Memorial prize in Economic Sciences in 2002[10] for their work applying psychological insights to the field of economics, showcasing how economic theory could not only be predictive, but also psychologically accurate. Some of their most relevant findings for the purpose of analyzing economic conception of agency are what they

[10] Amos Tversky unfortunately passed away in 1996 and the Nobel Prize is not awarded posthumously.

defined as prospect theory, which explores individuals' decision making under uncertainty. They actually devised this theory in exact opposition to the theory put forward by Von Neumann and Morgenstern.

Kahneman and Tversky regularly found in their experiments that individuals faced with uncertain prospects systematically deviated from the expectations of expected utility theory. The research began by with a seminal paper by Tversky titled *The Intransitivity of* Preferences. He claimed that "under specified experimental conditions, consistent and predictable intransitivities can be demonstrated" (Tversky 1969 p. 31). This led them to conduct research that they argue explicitly tested some of the axioms of rational choice theory (Kahneman and Tversky 1979). Their experiments consisted of questionnaires that were filled with hypothetical choice problems very similar to the ones that Von Neumann and Morgenstern proposed:

Problem 1:

Choice A:	2,500 with probability	.33	Choice B: 2,400 with certainty.
	2,500 with probability	.66	
	0 with probability	.01	
E(U) = u(2,500)(.33) + u(2,500)(.66) = <u>2,475</u>			*E(U) = u(2,400)(1) = 2,400*

Problem 2:

Choice C:	2,500 with probability	.33	Choice D: 2,400 with probability	.34
	0 with probability	.67	0 with probability	.66
E(U) = u(2,500)(.33) = <u>825</u>			*E(U) = u(2,400)(.34) = 816*	

I've added in the calculated expected utilities in italics below each problem. As outlined by Von Neumann and Morgenstern, when faced with a risky prospect, individuals should calculate their expected utility (E(U)) by multiplying the payoff by the probability of that payoff. I have also underlined the highest expected utility in each problem, which Choice A and Choice C being the two best choices for individuals. Surprisingly, 82 percent of respondents selected choice B in problem 1, and 83 percent selected choice C in problem 2. These results

systematically violate the prediction of expected utility theory. By selecting choice B in the first option, individuals are taking a lower expected utility for the certainty of getting 2,400. Yet selecting C in the second problem implies the reverse inequality, with the higher expected utility being selected despite the lower probability of it happening. Kahneman and Tversky take their analysis a step further to point out that the important violation is not only the fact that respondents didn't take the higher expected utility in problem 1, because maybe they were just extremely risk averse and thus avoided lower chances. Rather, respondents switched their preferences for risk between the two problems, despite answering an extremely similar question right after the first one with only a very slight probability advantage for one choice in problem 2. Their preferences were also not consistent between choices with seemingly similar payoffs.

Prospect theory seeks to explain this kind behavior that is difficult to account for within traditional rational choice theory models. Kahneman and Tversky saw that agents changed their preferences for risk based upon how the situation was framed to them. They called this the *framing effect*. Think about a time that you were in the supermarket going to buy yogurt. The container says "80% fat free" rather than "20% fat", and this is because empirical research tells us that individuals are more likely to buy the 80% fat free rather than 20% because of the way that the situation is framed to them. They also concluded that agents consistently weight outcomes that are certain much higher relative to outcomes that are probable. They called this effect the *certainty effect*. Individuals were much more likely to take certain gains over certain losses though, and the reason for this is because humans value gains and losses differently. People experience a greater emotional impact for a loss than they do for a gain of the same amount, driving their beliefs to be skewed towards being risk averse in situations that may not warrant it from a purely rational standpoint. All of this culminates in what Kahneman and

Tversky call prospect theory. Directly challenging the claims of Von Neumann and Morgenstern that agents faced with risky prospects will maximize their expected utility. Kahneman and Tversky explained, utilizing decades of empirical research, that agents under uncertainty were not only regularly unable to find the highest payoff, but also consistently held intransitive preferences, changed their preferences based upon framing of the situation and valued outcomes that were certain much higher in relation to risky prospects.

Their findings show that agent's decisions are not guided by merely rational evaluation and have formed the basis for their claims about how people ultimately make different decisions under uncertainty than in easy to understand, ideal situations that rational choice theorists examine. The intuition we get from this really is that the mathematical function doesn't fully explain the impact of the choices on people. This is extremely important because it counteracts the claim that positive economics (which just seeks to provide predictive/explanatory models) can still be useful despite its lack of psychological adequacy. We will look at the findings of laboratory studies which support this point in our next chapter, but it is important to see how these psychological inaccuracies are much more dangerous than theorists in favor of extreme abstraction may argue. Kahneman and Tversky's findings clearly show how actual human behavior does not line up with the assumptions of rationality in rational choice theory. They're not alone in their disagreement of these assumptions though.

Melioration

The psychologist R.J. Herrnstein contends, in explicit disagreement with the central claim of rational choice theory, that human agents do not maximize their payoffs, they meliorate (Herrnstein 1990). Melioration is the tendency to shift our behavior towards more lucrative alternatives, which likely provides higher reinforcement. This means that an agent's choices are

driven not by a maximizing function based upon the expected value of payoffs of alternatives, as Von Neumann and Morgenstern argue, but rather by choosing to shift our behavior towards the more reinforcing behavior. We are choosing to allocate our choices between different options based upon how the payoffs change as a function of how frequently we select an option. Herrnstein, along with other decision theorists, has run multiple experiments that show how individuals actually do not select the alternatives that lead to the maximizing outcome, but rather match the payoffs between outcomes by switching between them. Think about choosing between apples and bananas. You may like apples more than bananas all things considered, but when you've had an apple for the last 5 days in a row, you might choose to get a banana instead because there is a diminishing return associated with eating the apple.

There is an important time component in comparing maximization with melioration. When rational choice theorists claim that human beings are "utility maximizers" they implicitly (and sometimes explicitly) are referring to a decision made in a timeless vacuum, a specific point in time, a non-iterated decision. They are not accounting for the possibility of changing payoffs based upon previous choices. Now this type of abstraction is obviously useful because one can simplify the situation they are explaining (hopefully without losing significant predictive power). However, maximization does not provide that great prediction that is promised by proponents such as Milton Friedman.

Now, the time aspect of decision making wouldn't be a problem if the predictions that arose from maximization provided the same outcomes as melioration predictions, but it doesn't. Herrnstein's theory of melioration explains that our choices are affected by a variety of factors that move us towards conformity with what is defined as the matching law. Essentially, we attempt to distribute our behavior across alternatives in an attempt to equalize the payoffs

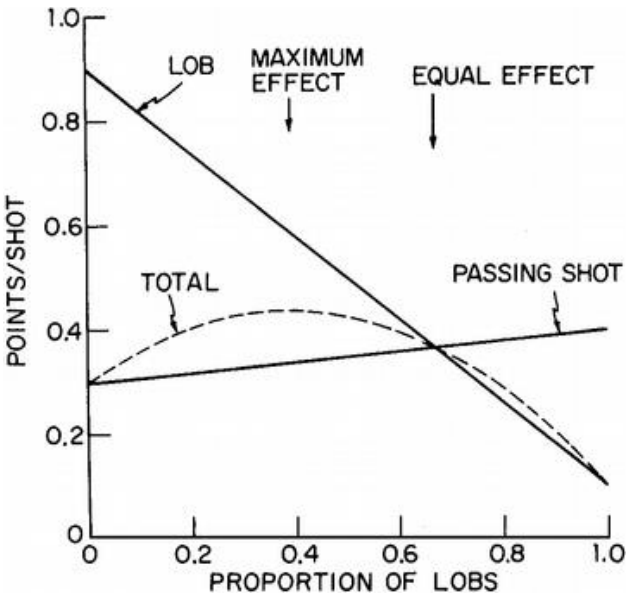
between each alternative. typically, is expressed by us switching our choices/strategies as one alternative gives a better payoff, but each time we make a choice, the payoffs change for our options.

Rational choice theorists may argue, and understandably so, that an individual is actually maximizing because in each individual decision, we are choosing the option with the highest payoff. But this fundamentally misconstrues the reason for why the person is making the decision. They are shifting their behavior towards more lucrative alternatives because the payoffs of all alternatives have been changing with each decision. Ultimately, the melioration that is done by the individual can actually lead to them making suboptimal choices and ending up with overall payoffs being worse than if they had used a maximizing strategy. The essential element not captured by rational choice theorists is the fact that choices do not happen in a timeless vacuum, they happen over time.

To better understand how these two views of human psychology produce different outcomes (and why one is more accurate), I think it's helpful to look at one of the examples that Herrnstein uses. Imagine that you are playing tennis, and your opponent chooses to approach the net after every time they return your shot. You have two possible choices in this scenario, either you could lob the ball over your opponent or try to hit a straight passing shot right by them. For the sake of the situation, we could say that a shot is more likely to be successful when it surprises your opponent, and each shot is equally likely to surprise them. Now, Herrnstein finds that most participants (and likely you) say that, given the situation laid out above, they would choose the shot that is most effective at surprising the opponent. This would be something like choosing to lob it until they catch on, and then start throwing in some passing shots and repeatedly switching back and forth as one shot does better than the other (switching behavior toward the more

lucrative alternative as payoffs change). Although this is likely exactly what we would all do in this situation, this would not necessarily be the maximizing outcome. In fact, it may be extremely suboptimal.

You can see on the graph that Herrnstein details below, melioration leads us to the point where the two solid lines intersect, and the points per shot equal each other. A maximizer would instead calculate that the ideal basket between passing and lob shots is at the highest point on the dotted line, where they score the most possible points per shot. This is somewhere around 40% lobs and 60% passing shots, significantly different from the equal effect point where you lob 65% of the time and pass 35%.



By switching between different choices based upon their perceived payoff (melioration), we would eventually end up with payoffs between lob and passing that are very close to each other. However, if the probability of a lob shot being successful when it was a surprise was much higher than a surprising passing shot, then the maximizing point would actually be a combination of lobs and passing shots that is much different from our point where the payoffs between the

two choices match. You would overcompensate by repeatedly choosing to lob, because the reinforcement for choosing the option would be much higher.

Maximizing and Meliorating create 2 different outcomes in this scenario, an "optimal" strategy, and an "equilibrium" strategy, respectively. Psychologists in favor of melioration argue that we do not actually achieve the optimal strategy, but rather one that is 'good enough'. This creates problems for theorists' prediction because the mathematical model predicts a maximized outcome, but the real agents select a different equilibrium. This creates further problems for rational choice theory because every single one of their axioms of rationality suffer to be accurate when tested against actual human behavior. There are some economists and psychologists who recognize these kinds of psychological limitations on agents though and seek to argue for an adapted conception of rationality in rational choice theory.

Bounded Rationality

Herbert Simon, a psychologist who has made large contributions to the field of economics, argues for a much more relaxed conception of rational choice theory called bounded rationality. He defines it best himself: "The term 'bounded rationality' is used to designate rational choice that takes into account the cognitive limitations of the decision-maker - limitations of both knowledge and computational capacity" (Simon 1990 p. 15). Simon asserts that a lot of the issues in prediction that arise for rational choice theory are due to the limited cognitive capacity and available information to human agents. Rather than scrapping rational choice theory altogether, he proposes that we should relax some of the core assumptions in order to better represent the decision-making process.

Chief among these considerations for revision are humans' ability to generate possible alternatives, being consistent about their choices and the availability of information. Studies of

human decision-making show clearly how agents exert a great deal of effort in coming up with possible alternatives, rather than experiencing them as givens like rational choice theory assumes. This is really important because it puts a huge hole in the first axiom of rational choice theory. Agents are not actually aware of all their possible alternatives when faced with a decision, they must expend a significant amount of mental effort just to discover some of them let alone all of them. This is a big reason why we regularly select a choice that is merely satisfactory rather than 'best' or why we may make choices that are suboptimal because they are safer and require less calculation (Simon 1955).

Humans experience cognitive limitations in computational power. This is especially true in economics where the accuracy of an agent's predictions about the economy is inherently limited by their understanding of the underlying economic mechanisms. Most consumers have very little knowledge about how market systems work, or any of the other underlying economic mechanisms that may help them to make a more informed decision. For Simon, bounded rationality helps to account for these kinds of limits on human decision making by making predictions based upon a relaxed ideal of rationality, one in which humans try to achieve their goals by using their minds as well as they possibly can given their cognitive constraints. This view of decision making also grants much more explanatory power when faced with an agent making decisions that are not optimal. Simon shows us that there is a path to recognizing the psychological reality of agents in the traditional models of rational choice.

Handling psychological shortcomings

Let's look back at the four axioms that we originally attributed to rational choice theory as describing the economic agent: they have a known set of alternatives; have preferences about each alternative; their preferences are transitive; and they maximize their payoffs. As we

compare these axioms to the findings from the psychologists that we discussed above, it's clear that there are massive holes within the 'economic man'. Simon shows us that human beings struggle with being able to generate all of their possible alternatives, casting doubt on the assumption that agents have a known set of alternatives. Kahneman and Tversky found that framing effects are extremely important on decision making and that our preferences are not actually consistent when faced with similar choices. Finally, Herrnstein provides a fabulous account of the nuances of melioration, and why it is a much better theory of decision making than maximization. The central axioms of rational choice theory suffer from serious psychological concerns and cast doubt on the theory's ability to accurately describe agency and predict future action.

In previous chapters, I discussed how rational choice theorists argue that reducing human agency down to rationality and behavior reduces the number of inputs needed to build effective models. But agency becomes a thin shell, describing nothing more than a web of preferences and a methodical calculation to determine which actions align most with our preferences. There is no discussion of mental states, such as beliefs or desires, or any kind of commitments to principles, such as morality or fairness. The mainstream viewpoint following the creation of expected utility theory in the 1940s was that all of complex human agency could be reduced down to a mathematical function, a utility model. The descriptive psychological accuracy of this theory is clearly low; however, that may not matter if we are able to genuinely predict how agents act in an economic setting because they are acting *as if* they are maximizing their expected utility. The psychological findings that we discussed here expertly point out how complicated and how limited human cognition is when faced with complex decisions. The work done by Kahneman and Tversky was imperative for the beginning of behavioral economics and opened the window

to exploring decision making beyond the limits of rational choice theory. Their work showed how psychology is not only relevant but necessary for furthering the field of economics, and thus directly addressed the claims of economists who said that psychological abstraction was necessary for reasonable prediction and explanation.

Results from Behavioral Economics

Behavioral economists seek to provide a more psychologically accurate picture of human decision making. These economists test the limitations and strengths of the assumptions about agency made in rational choice and game theory by conducting experiments that test the actual behavior of agents. The field developed out of the work done by psychologists Daniel Kahneman and Amos Tversky in applying their psychological findings in laboratory studies to the questions posed in economics. Although done so without intention of revolutionizing the way we think about economics, they formed the foundation of a new school of economics which better accounts for the wide range of human psychological possibilities. Their work studying framing effects on people's decision making, showed the world how an individual's environment, especially the way they are presented a problem, is a major factor in their final decision making. In the late 1970s, Kahneman and Tversky worked alongside an economist by the name of Richard Thaler, who helped to bring some of the challenges that faced classical utility theory in line with the findings of psychology. Thaler ultimately was awarded the Nobel Memorial Prize in Economic Sciences in 2017 for his work in furthering the field of behavioral economics. His findings included the now well known 'endowment effect', which is when people value the same item more highly when they own it as opposed to when they are deciding whether or not to buy it (Thaler 1980) [11].

[11] He also defined the term 'nudging' as a method to encourage consumers to make better choices for themselves through subtle yet effective methods, such as putting fruit at eye level in the store or even putting an image of a fly at the bottom of a men's urinal (Thaler & Sunstein 2008). Thaler wrote an outstanding book on the topic which you could read for more information; but the main idea behind it is that people don't always respond well to explicit controls to their environment. Instead, using small changes that can 'nudge' them in the right direction is a very powerful method to change people's behavior for the better. It is certainly up for debate whether or not we should be intentionally influencing people's behavior or if that takes away their autonomy.

As evidenced by Thaler, behavioral economics has begun to be recognized as a powerful aspect of the study of economics. Thaler's research, similar to other behavioral economists', emphasizes recognizing the psychological reality of human agents when studying their economic behavior. This is important because, as he discusses, we build our political policies and organizations based upon how we believe human agency functions (Thaler 2015). If our conception of agency is incorrect or misled, then we could create incentives that are ineffective or push individuals in the wrong direction. We could be encouraging malicious or selfish behavior incidentally, rather than trying to build a society focused on the values that we care about most (whatever they may be). Behavioral economics helps us to discern which factors in our world are significantly impactful for human agency, and therefore helps us see the limitations of using rational choice theory to predict economic action.

Behavioral economists do this through a variety of laboratory games which help to reveal an individual's preferences or specific influences on their decision making. These games often come from game theory and the experimenters test whether or not the theory's predictions are actually played out in the laboratory. The reason this is so relevant for purposes in this paper is because many of the results of behavioral economics studies have actually shown the initial predictions by traditional rational choice theory about human behavior to be incorrect, or at least failing to capture the full picture. Exploring these results is crucial if we want to figure out the place that rational choice has within economics' conception of agency and how it can be realigned with competing claims of agency. I will start by looking at two basic games, and then move on to 2 more complex games. After describing these games, predictions about the outcomes, the results that were revealed, and interpretations of the results, then we will look at

the potential responses from rational choice theorists as well as the impact that these results have on the overall robustness of the predictability of traditional rational choice theory.

Dictator and Ultimatum Games

For all of these games, imagine that there is a total pool of $10 being used, but you could imagine that any amount of starting money is used. A dictator game is one in which there are two players, a dictator and a receiver. The dictator is given $10 and is asked how much he would like to share with the reciever. However much the dictator chooses is then given to the responder and the game is over. An ultimatum game is a variant of a dictator game which adds the possibility of response. There are two players but no dictator, just a proposer and a responder. Once again, the proposer is provided with $10 and asked how much they would like to share with the responder. The proposer selects an amount and that is given to the responder. The responder is then given the choice whether to accept that amount or reject it. The ultimatum aspect of the game is that if the responder chooses to reject the money, then neither player gets anything. So, this game has a *strategic element* involved which the dictator game does not, as the decisions made by the proposer does not unilaterally determine the outcome. Instead, the proposer must be cognizant of the responder's choices because they affect the outcome of the game. There are numerous variations of these games, including differing levels of available information for each player, such as telling the responder the money is from a 'pie in the sky' or telling the proposer they are a dictator rather than a responder, and limiting the choices of how a proposer/dictator can split up the money (such as choosing between an 8/2 split or 5/5 split rather than any amount they want).

Traditional rational choice theory tells us that players in these games should approach the situation by mapping out their possible outcomes, determining the one that gives them the highest payoff based upon what their preferences are within the given game, and then maximize

by selecting that payoff. Now when looking at participants playing these simple games, it is not immediately clear what their preferences are for choosing how much money they want. As we discussed earlier though, there is a consistent assumption within rational choice theory that given a simple choice between more money ($10) or less money ($0), any economic agent would choose more money. Following from this, dictators, in a one-shot iteration of the game, should give the smallest amount possible to the receiver because there is no chance of backlash from the receiver. Their maximizing, rational choice is to take as much money as they can because there are no strings attached (the receiver cannot reject the offer). I will acknowledge here that this is also assuming that individuals do not have any consideration for fairness or equal outcomes though, which is a consideration we will discuss later on[12].

For the ultimatum game, the proposer has to consider the other player's desires because they can reject the proposal if it is too low. In this situation, even if it is a single shot game and there would be no future repercussions, there are strings attached to how much money the proposer can walk away with. Nevertheless, it isn't exactly clear how much the proposer should propose, as it depends on the preferences of the responder. However, on the flip side, responders in the ultimatum game should theoretically accept any money that they are offered because it would make them better off than they were before. If we follow game theory's rationality assumption that the proposer is assuming that the responder will select their dominant strategy, it makes sense for the proposer to propose as little as they are able. Even if the proposer offers $1 to the responder, the dominant strategy for them, given the payoff of $1 vs nothing, is to take the dollar.

[12] Some modern game theorists/ rational choice theorists have attempted to create models that actually account for these types of preferences within a traditional utility function.

Now, all of these predictions are preliminary and most come from the utility theory developed by Morgenstern and Von Neumann in the early 1940s. Since the results of behavioral economics, there has been significant work done by game theorists and rational choice theorists to incorporate broader preferences than merely 'more is better than less'. We will discuss these considerations as well, but I first want to look at the results of the dictator and ultimatum games and why they were surprising in light of traditional rational choice theory. The findings of behavioral economics studies do not support the type of behavior predicted by the traditional considerations of rational choice. Colin Camerer's book *Behavioral Game Theory: Experiments in Strategic Interaction*, details many of these anomalies and generalizes the results of a large body of games conducted. These are the findings from ultimatum games:

> *"Modal and Median ultimatum offers are usually 40-50 percent and means are 30-40 percent. There are hardly any offers in the outlying categories of 0, 1-10, and the hyper-fair category 51-100. Offers of 40-50 percent are rarely rejected. Offers below 20 percent or so are rejected about half the time" (Camerer 2003 p. 49).*

The important takeaway from the results of the ultimatum game is that people consistently give a much larger amount of their initial pool than traditional rational choice predicted they would. Now, in ultimatum games, there is risk of retaliation from responders if the proposer offers too low of an amount. So, it could be argued that proposers are choosing to give some money in order to ensure that the responder doesn't reject their offer. There is no doubt that there is some element of this going on, as it is a fundamentally strategic game, and the proposer must appeal to the responder in order to get the outcome they want. The predictions from game theory, which argue for proposers giving the smallest amount possible and the responders accepting any money, are not consistent with the results. At minimum, the results cast doubt on the consistency of this type of rational calculation being done by agents in these settings. The

offers are consistently more than just a small amount (30-40 percent on average), as if the proposers are recognizing the desires and humanity of the of the responder.

However, on the other side, the responder also chooses to reject offers half the time when they are around 20 percent of the total or lower, implying that they have some type of idea of what is 'fair' to be given to them. This commitment to fairness is completely inexplicable in the model of rational choice theory. Agents who participated in the ultimatum game consistently chose to take a lower payoff for themselves in order to punish the proposer who offered an unfair allocation. This is very puzzling because it is not an appeal to altruism or any kind of self-interested preference, but rather an appeal to a normative principle – the way the money ought to be allocated.

I want to clarify what I mean here by agents caring about *fairness* rather than an *altruistic motivation* to share the pool with the other participant. Authors use these terms in a wide variety of ways, but I think it's important to distinguish between the two especially for our purposes. The key difference here is in between a feeling of sympathy for others and a commitment to a principle. The concept of fairness is a principle that people can be committed to and shape their action in accordance with. Altruism is a motivation to help other people because you are concerned about the welfare of the other person. It is a desire to help other people and it motivates them to act in favor of others' desires. Daniel Batson runs some great experiments and argues for how empathy can evoke altruistic motivations in individuals (Batson 2010). Altruism can be explained through rational choice theory as merely a higher utility payoff for an agent. However, commitment to the principle of fairness could be completely counter preferential.

This is an important distinction because when we look at the results of the ultimatum game, choosing to share the pot with someone because you actually desire them to have some of

the money is a dramatically different from choosing to reject an offer, and therefore decreasing your own payoff in order to punish an unfair outcome. Someone who is altruistically motivated is doing it for the sake of the other, rather than for the sake of a principle, and could still be maximizing their payoffs in accordance with their preferences because they actually *prefer* to give money to the other person (to a certain extent). This is an interesting distinction because it separates out different motivations that an agent could have for making a choice. In fact, you could even imagine a scenario where someone is committed to a principle and committed to other people in a conflicting way. You want to be fair to two individuals and split money between them equally, but one of them is your friend and you are more sympathetic towards his needs than towards the strangers. These two motivations are not reducible into merely preferences, but rather appeal to an interesting constraint on agents – normative commitments.

Rational choice theory seeks to reduce everything down to 'preferences' when it comes to explanations for human motivation and choices. However, the differences between an action motivated by fairness versus an altruistic action are important. By being able to descriptively talk about agents in a more complex and holistic way, we are able to better determine why certain predictions don't come true in empirical studies. We lose subtlety when we reduce everything down to preferences, which is why becoming more psychologically accurate is imperative for the theory to provide more predictive accuracy. This distinction is more clearly defined when we compare the studies of dictator games with ultimatum games.

Within dictator games that Camerer and others conducted, the mean allocation of money when proposers could choose to give any amount of money, they want was about 20 percent (Camerer 2003). This number is much smaller than the findings of ultimatum games, which tells us that there are different motivations or principles being evoked between the two games.

Camerer interprets this to be evidence of some form of 'pure altruism' influencing dictators' decision to give more than nothing because they should not have any fear of retaliation (Camerer 2003 p. 56). This is obviously a very complex discussion in terms of interpreting these discrepancies, but I argue that these results certainly point to some kind of influence on individuals to give more than nothing. However, that does not necessarily mean it was altruism that drove them to act in that way. Herein lies the difficulty of interpreting the results of behavioral economics, because it is hard to say which specific factor it was that led to proposers sharing a larger amount than rational choice theory suggests they should in dictator games. It may be that they are merely trying to appear as fair to the other person rather than actually caring about the outcome for the responder. I do think that these results provide some clear evidence for the existence of outside influences beyond self-interested motivation though, which may have been originally overlooked by game theory and rational choice theory.

These results parse out a distinction between the factors evoked by ultimatum game versus dictator games. Dictators do not need to necessarily care about fairness, as there is no enforcement mechanism against 'unfair' actions. Violations of the principle cannot be sanctioned in any way. I think it is a more plausible interpretation that the dictators are acting out of some altruistic motivation to recognize that the other person needs some money too which may be informed by a principle of fairness or kindness. However, the ultimatum game lets us focus explicitly on considerations of fairness rather than altruism by looking to the responses of the responder. The responder only has two possible moves – accept the money or reject it. According to empirical results, offers that are below 20 percent of the total pool are rejected half of the time, which means that half the participants are choosing to have no money over some money.

As I mentioned, this choice is extremely puzzling within the context of rational choice theory, especially since there cannot be any preference for altruism evoked here. If we look back to expected utility theory – the agent in this context is choosing to take the lower utility option in one situation (nothing rather than a small amount), but then is not offering similar amounts when they are the dictator. Their preferences do not remain consistent, *or* they are making a counter-preferential choice when they are a responder in the ultimatum game in order to punish others [13].

Some researchers have argued that by utilizing ultimatum games and dictator games we can establish a baseline for an individual's preferences across games and they can then be tested in other, more complex games (Guala 2006). The idea behind this is that by viewing an individual in a simple game, we can see what their preference is for fairness or altruism. Once this has been established, you can then put the player in a game that is similar but involves some more complex element, such as reciprocity, and see if preferences remain consistent. Guala argues that experiments have now been run that test for this explicitly and so we can see if there are genuine implications of reciprocation. This is what brings us to looking at some more complex games within behavioral economics. For example, we can observe the amount shared by an individual in a dictator game, and then place them as the responder in what is called the investment game. The findings of these experiments show that behavior is different for an individual in the dictator game vs. the investment game. This means that the investment game provides an added influence on their decision making that *is not present in a non-strategic dictator game*.

13

Investment and Public Goods Games

The investment game is also with two players, an investor and responder, but adds elements of trust and reciprocity that is not present in the other games. For the investment game, the investor chooses how much of the $10 he wants to give to the other player. Then, the experimenter takes the money that was given (let's pretend it was $5) and triples it before giving it to the responder. In our example, the responder would receive $15. Then he is asked how much of the $15 he received he would like to give back to the investor. As you can see, he could give $5 back in this situation, and then they would both end up with $10. This game adds an element of trust between the two players because not only is there reciprocation, but the responder could have the opportunity to keep more money than they were originally given. The investor must trust that the responder would be willing to give back some of money that they 'invest' in them.

The public goods game typically has more than two players and is centered around donating to a goal that is good for everyone (a public good). These games are often modeled by giving everyone a certain amount of money to start with, let's say $5, and then telling them that however much everyone donates to the middle pot will be doubled and divided evenly amongst everyone. The experimenter then asks each participant how much they would like to donate to the public investment. There is a well-documented phenomenon within economics surrounding public goods called the free rider effect, in which individuals are incentivized to not contribute to the public good because they will be able to benefit from it no matter whether they donate or not (such as deciding whether to donate to a public park). The public goods game tests for considerations such as the free rider effect and also how the ability to punish free riders can impact contributions.

As I just mentioned, some game theory researchers, such as James Cox and Francesca Guala, have argued that we can empirically test for rational behavior of agents by combining experiments of simple games and more complex games. The crux of the argument behind rational choice theory is that if you are able to determine someone's preferences, you could draw up a utility function and determine which options give them their highest utility (payoff), which will then be the option they would choose. If people's preferences for equality, reciprocity, and/or altruism could be revealed in a non-strategic setting, then they could be tested again in a more complex strategic setting to see if they hold up and are rationally consistent. James Cox acknowledges that traditional game theory assumptions constrict agents to only having self-regarding preferences (Cox 2004). This means that any and all choices must in some way relate back to themselves or some gain they are going to receive. This ignores the results of behavioral economics which showcase evidence for altruistic and reciprocal motivations. Let's first look at his experiments to see how game theory can work to still account for these empirical findings.

Cox conducted experiments in which he utilized a traditional investment game and a dictator variation of an investment game (Cox 2004). There were two players, player A and player B, and the experimenters conducted three different types of tests. In the first test, it is just a simple investment game with player A being given $10 and choosing how much to give to B. That money is tripled and given to B, and then B chooses how much to give back to A. B is fully aware that A was the one who gave them the money and A knows that the money is going to be tripled, so the experimenters anticipate that feelings of reciprocity are present for B and feelings of trust (that they will get some money back) will be elicited in A. The second variation is a dictator game in which A chooses how much to give B, that money is tripled but then B cannot give any money back to A. This variation tests for 'other regarding preferences' such as altruism,

without the motivation of trust for A that they would get something in return. Finally, the experimenters utilized a random selection of distribution that one of the A players made in the first test and presented them to B as a starting pool (as if they were the A player) and asked how much they would like to give to A. This test removes any feelings of reciprocity from B, as they are no longer motivated by A giving them that amount of money. Cox argues that these three tests separate out motivations from altruism, trust and reciprocity for individuals.

The results that came out of Cox's experiments showed that over half of A players (63%) chose to give *positive amounts* (more than 0) of money to B in the dictator variation. Based upon the controls of the experiment, these results can be interpreted as motivation from the other-oriented preference, altruism. When these results are compared to the first test, a traditional investment game, even more of the A players chose to give positive amounts to player B (80%). Additionally, significantly more responders chose to give the maximum amount of $10, (13 responders versus 4 in the dictator variation). These results point out that there are genuine differing motivations going on when people are given the chance to perform a purely altruistic action, as opposed to a trust action. There is clearly a complex interaction going on between selfish desires, other-oriented preferences and commitments towards other people. The traditional game theory and rational choice models struggle to incorporate these kinds of concerns into a utility model, but once preliminary results from behavioral economics began to surface, game theorists began to try to incorporate them. These results have led to new utility models being created by psychologists and behavioral economists that try to incorporate these abnormal influences on decision-making.

One of the more popular models of this concept is called the Fehr-Schmidt Model, coming from their research done in the late 1990s (Fehr & Schmidt 1999). Similar to Guala, this

model proposes that we can identify a subject's utility function by studying their decisions in one-shot games (such as dictator or ultimatum). This then would allow us to predict a subject's behavior in other games with small variations. Furthermore, Feher and Schmidt argue for the existence of stable preferences of individuals, keeping consistent with game theory assumptions, which allows them to predict how someone would act later. There is a much more ambitious and holistic model (in my opinion) created by Matthew Rabin, who argues that agents care not just merely about the outcomes of the game (which is what the Fehr-Schmidt model argues for with inequality aversion), but also about the intentions and motivations of the player acting. Essentially, he argues that "people like to help those who are helping them, and to hurt those who are hurting them" (Rabin 1993 p. 1281). We don't care as much about the actual outcome of the game or transaction, but about the way that we were treated by the other individual. He uses evidence from public good and ultimatum experiments to show how individuals choosing to donate to a public good is dependent on the choices made by other agents within the game.

Although the standard rational choice model would suggest that everyone should simply be a free rider in a public goods game and avoid contributing to the public good, evidence shows that people contribute to the public good at a much higher rate than expected. In fact, the contribution rate for these games range between 40 and 60 percent across a large body of experiments. In further support of Rabin's theory, there is a substantial body of psychological literature related to punishing others. Individuals are more likely to sanction unfair actions and protest that action than they are to sanction themselves when they are presented with a better, but unfair option (Fehr & Gachter 2002). Coupling this with evidence from ultimatum games, it seems evident that agents are actually willing to sacrifice their own gains in order to punish someone else.

The reason that I bring these models up is to showcase that there are some economists seeking to adapt the traditional economic utility model in order to better account for the rapidly growing database of research in behavioral economics. There is certainly much more work to be done but it shows that there are pathways to keeping the core of rational choice theory whilst recognizing other influences on decision-making. The results that come out of behavioral economics and these models that try to incorporate fairness/ inequality aversion within them and provide evidence for influences on agency that are beyond merely self-interested preferences. These considerations could be included within traditional utility functions, but it is difficult to say that normative commitments, such as fairness, could.

Ultimately, I think that the clearest takeaway from behavioral economics is that experimental results, and the theories of rational choice theory both struggle to have their predictions extrapolated outside of very specific, constrained environments. Rational choice theory is most predictive when the preferences of an individual can be known, their environment is constrained, and there are limited possibilities for an agent to identify. We saw this with the FCC spectrum auction when theory was able to tell us which rules of the auction to be cognizant of, but we needed to do specific testing in order to figure out the exact influence that a rule would have on the outcome of the auction.

This means that one of the largest limitations on both rational choice and game theory for images of agency is that they only provide answers in specific limited contexts, but do not do well when being generalized to a paradigm of agency and decision making. This is especially true when commitments to principles are evoked, as the specific principle can cause individuals to change their behavior away from merely self-oriented preferences. It is fairly clear, by the evidence that experimental economics has provided, that human agents are not purely motivated

by selfish desires and utility maximization, but also influences including other-oriented preferences and commitments to principles.

There are some rational choice theorists who contend that these influences on decision-making can all be accounted for through selfish preferences. For example, Ken Binmore contends that when agents play one-shot games in an experiment, they are just importing tactics that they would use in repeated games because all of their life experiences are generally repeated strategic situations (Binmore 2007). He concludes that it's actually merely excessive to appeal to the existence of any factor beyond selfish preferences when explaining one-shot games. This is because individuals are not acting out of kindness/ altruistic motivation toward the other player but are merely importing the tactics that they used in repeated games (such as tit-for-tat strategies). Woodward cleverly identifies the bigger problem that this shows for all of behavioral economics, which I hinted at previously, which is that the experimenters have weak control over not only the type of game that is being played but also the type of behavior that is being elicited (Woodward 2009). It's not immediately clear whether individuals are being altruistic, inequality averse, or acting out of commitment to a principle when they play these games.

This discussion about whether people are inputting strategies from other games is a broad debate which I don't have space to fully discuss here but will note some quality articles that make good sense of it[14]. However, I will say that there is significant evidence based upon questions asked before and after experiments that clarifies that participants understand the game that they are playing; and, arguably more importantly, there is evidence of participants choosing to play games differently when they in are one-shot games versus repeated games (multiple rounds with the same partner) (Camerer & Fehr 2004). Rational choice theorists such as Binmore

[14] Binmore 2007, *Game Theory: A very short introduction*; Samuelson 2005, *Economic Theory and Experimental Economics;* Henrich et al. 2004, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*

refuse to address the clear evidence that normative influences genuinely effect that way that agents make decisions.

These influences not being considered in the paradigm of agency in economics has been a path that economics has taken in its quest to heavily mathematize the discipline and limit appealing to non-verifiable states. Some philosophers have been questioning for a while whether the classical economic model of agency accounts for what they regard as the embedded social aspect of human agency, such as Adam Smith and Amartya Sen. I have argued so far that the model does not fully captures the wide range of commitments and influences that human agents are beholden too. However, it also focuses so heavily on the individual that it misses the intrinsic social nature that moral sentimentalist philosophers see in all human agents.

The reason I bring up this whole discussion is because it does lead us to a very plausible response to the results of behavioral economics by rational choice theorists, and an answer that many of my own economic professors have given at one time or another, which is that when someone chooses an option with a lower payoff when faced with different alternatives, then we merely 'got people's preferences wrong'. When rational choice theorists predicted that someone was going to act selfishly and choose to give nothing to the other person in a dictator game, they just mistook his preferences and did not account for his preference for altruism. In this sense, there is no such thing as a 'counter-preferential' choice. Every intentional action is a maximizing one under this paradigm of agency. Amartya Sen, a Nobel prize winning economist and philosopher, recognized this point in his article critiquing mainstream economic theory, *Rational Fools,* back in 1977:

> *It is possible to define a person's interests in such a way that no matter what he does he can be seen to be furthering his own interests in every isolated act of choice. While formalized relatively recently in the context of the theory of revealed preference, this approach is of respectable antiquity, and Joseph*

Butler was already arguing against it in the Rolls Chapel two and a half centuries ago. The reduction of man to a self-seeking animal depends in this approach on careful definition. If you are observed to choose x rejecting y, you are declared to have "revealed" a preference for x over y. Your personal utility is then defined as simply a numerical representation of this "preference," assigning a higher utility to a "preferred" alternative. With this set of definitions, you can hardly escape maximizing your own utility, except through inconsistency (Sen 1977 p. 322).

Sen was not fully aware at the time of writing this article of the bounds that economists would take to try to account for preferences outside of the traditional utility model. Yet, his criticism still rings true to the problem faced by making choice synonymous with preference in a theory of agency – it creates a circular, self-fulfilling prediction. If an agent is faced with a choice between an apple and a banana and he chooses a banana, then he reveals his preference of bananas over apples; and they would predict that he would make the same decision next time he is faced with that choice. However, if he chooses an apple over a banana next time, the observer can just readjust the attributed preferences rather than admit that he selected a counter-preferential choice or that his preferences are inconsistent. Every single time an agent makes a choice that is different from his previous, rational choice theorists can remake his utility model so that his new choice is still the maximizing option. His behavior is being explained by his behavior, which tells us virtually nothing. If the response to the results of behavioral economics findings is merely "we can just readjust people's preferences" then you dramatically reduce the predictability and explanatory power of rational choice theory.

As you may be able to tell, Sen has almost no faith in the power of rational choice theory because of its limited conception of human agency. He attributes much more of an emphasis on agents being selfish to economics than I would personally, but he points out significant problems with the theory which I'd like to address in relation to the results of behavioral economics. He views the 'economic man' as a selfish individual with zero social awareness. I would contend that mainstream economics does account for other-oriented preferences and not merely selfish

ones but agree that there is little to no consideration for commitment to principles such as social norms. When Sen wrote his article in the late 70s, he was without the significant empirical evidence that has come in the last 2 decades of research in experimental economics. He made claims made about rational choice theory's paradigm of human agency that I argue are now supported by the results we just looked at. Sen defines a separation between feelings of 'sympathy' and 'commitment'. For Sen, sympathy is when we are made uncomfortable by the suffering of others, and thus are egotistically motivated to alleviate their suffering in order to stop our own discomfort[15]. These choices can be fully consistent with utility maximization as it makes the agent better off by helping the person. Commitment, for Sen, is an act out of duty to a principle and is regularly counter preferential. That is, an agent chooses to have less personal welfare in order to keep their actions in line with a principle to which thei are committed.

Sen uses a very simple example which I think is worth exploring here:

"The contrast between sympathy and commitment may be illustrated with the story of two boys who find two apples, one large, one small. Boy A tells boy B, "You choose." B immediately picks the larger apple. A is upset and permits himself the remark that this was grossly unfair. "Why?" asks B. "Which one would you have chosen, if you were to choose rather than me?" "The smaller one, of course," A replies. B is now triumphant: 'Then what are you complaining about? That's the one you've got!'" (Sen 1977 p. 328)

This seemingly silly example showcases an important distinction here between the two types of influence on decision-making. Boy A had an expectation that friends who are committed to the norm of friendship choose a smaller apple when faced with the choice between a larger or smaller apple and therefore Boy B ought to choose the smaller one. His frustration when Boy B selects the larger apple comes from the violation of this perceived normative commitment to friendship. Boy B clearly has no such expectation and therefore chooses the larger apple,

[15] Sympathy is contemporarily known as the genuine concern for the welfare of others. What Sen is describing is more like personal distress.

maximizing his personal utility. Boy A's response that he 'would have chosen the smaller one' is not because it was his preferred option, but because that is what friends *ought* to do. Sen is tapping into the very complicated decision process that agents engage in when faced with commitments such as social norms, and now I would argue there's empirical evidence to support his claim. Rational choice theory does not help us to explain why Boy B is frustrated, because if the smaller apple was his preference, he has no reason to be upset as he is maximizing his utility.

Reducing this situation down to merely preferences does not explain why Boy A is so upset at Boy B. Boy A is frustrated not because he preferred to get the bigger apple, he's totally okay with receiving the smaller apple, but because Boy B did not act like a good friend. This type of frustration cannot be explained by preferences and can lead to faulty predictions if it is not properly accounted for. I argue that Sen's intuition, that commitments actually change people's preferences and can make them choose worse payoffs, is supported by the fact that they choose to make themselves worse off when they are influenced by normative commitments in laboratory experiments. There are some philosopher-economists (really just one), who have been working recently to incorporate social norms into the rational choice model.

Bicchieri's Social Norms

Really great work has been done by Cristina Bicchieri in trying to close the gap between rational choice theory and the findings of behavioral economics when it comes to social norms. Bicchieri is firmly within the rational choice side of the debate but works as a philosopher-economist to account for social normativity. First let's understand quickly how she believes a norm exists. She has 3 key criteria. There must be a sufficiently large subset of people who are aware of a behavioral rule R in a group (Sufficiently large changes based on the group). This

subset will prefer to conform to rule R, condition on their belief that others conform, and that they believe that sufficiently large subset expects conformity (normative expectations); or believe that people expect conformity and may sanction behavior that does not conform (normative expectations with sanctions). She keys into the fact that norms are very often sensitive to contextual factors like framing and intentions and that different contexts or frames trigger different norms in otherwise similar games.

Her experiments with Chavez worked on measuring people's normative expectations about what a 'fair outcome' was in an ultimatum game and then testing how that affected their decision making (Bicchieri & Chavez 2010). Bicchieri would not necessarily agree that these influences are 'non-selfish' in nature but would rather merely describe them as normative influences. She claims that people have expectations of how others ought to act, and that in turn influences their judgement of others' actions and their own action. Bicchieri argues that experimental evidence from ultimatum games shows that there is a role played by expectations in influencing an agent's behavior in social settings (Bicchieri et al 2018). I certainly agree with her, but it is interesting to see how she works to include this within the model of rational choice.

The experiment gave participants a choice between splitting $10 either in an ($8/$2) split, ($5/$5) split or let a coin decide between the two. They then measured people's normative expectations of which options were 'fair' by having them fill out a questionnaire, noting that there was a significant amount of agreement as to which options were fair (most agreeing that ($5/$5) and (Coin) were both fair choices). They then conducted ultimatum games with the participants giving them varied levels of information (full, private, and limited). Private meant that they did not know whether or not the coin option was available, and proposers were aware that responders didn't know; limited allowed responders to know coin was available, but they did

not know whether or not it was selected, just the final allocation. Their results found that when normative expectations were not present in the experiment (because the responder didn't know coin was available or couldn't tell the difference) then the proposer chose a selection that they preferred much more frequently than the (Coin) option. They were certainly influenced by what they thought was the 'fair' outcome but selected both ($5/$5) and ($8/$2) much more often than (Coin). The selection of (Coin) was highest in the full information condition, presumably because the normative expectations were present, and the responder would be aware of the 'unfair' selection by the proposer.

Bicchieri argues that this is because the social norm is evoked in the proposer and thus constrains their action once the responder is aware of which choices are made available to the proposer. The context of the experiment is extremely important in determining which outcome occurs, and the payoff for each individual is significantly influenced by the activation of the social norm. Bicchieri bridges the gap between rational choice theorists who contend that only self-interested preferences exist or that you can reduce everything down to a preference and philosophers and psychologists who interpret these results as saying that there are normative influences in addition to someone's desires.

Normative pressures could be a powerful explanation for why agents do not abide by the predictions of rational choice and game theorists. For Bicchieri, the cooperation outcomes for strategic situations are actually a part of the utility function through social norms. For Binmore and traditional rational choice theorists, a norm is simply a Nash equilibrium in a game. It occurs because people have certain preferences, and the game is structured in a specific way. The norm is not a preference in itself, it does not enter into the utility function. Bicchieri views norms as preferences though and argues that agents have preferences for conforming to different norms

and that violation of this preference to conform actually damages their utility function. The reason this is so ingenious is because it addresses the concern that Sen raises about commitments. Bicchieri's utility model recognizes that an agent's *personal welfare* could be worse by choosing to conform to a social norm, but their *utility* would still be higher. Her model expands the definition of utility to involve broader concerns than merely an agent's own personal welfare.

Bicchieri's model is absolutely a step in the right direction of working to incorporate normative commitments such as social norms into the model of utility functions. However, as I discussed earlier with Sen, there is still a serious explanatory problem with reducing the varying motivations and contextual influences on an agent down to merely their preferences. I argue that the psychological inaccuracy of rational choice theory is actually leading to poor predictive capabilities for the theory as a whole, even if you work to incorporate social norms as a preference. It suffers from the same type of issues of circularity. However, this does move us closer to understanding agency in relation to normative commitments and social influence. Adam Smith, the father of classical economics, saw human agency as strongly influenced and even constituted by our feelings of empathy for one another. Sentimentalist philosophers' intuitions about human agency have found further support in the empirical studies that we just looked at, which suggest that human rationality, as mainstream economics sees it, should be reconceived in a way that recognizes adherence to social normativity as more than something that we choose merely for selfish reasons. Rather, sociality may be more directly ingrained into the basic structure of thinking about the world in a rational manner.

Smithian Economics

As we've seen through the results of experimental economics, the paradigm of agency within economics, the economic man, struggles to accurately predict individual's decision making outside of constrained, specific situations. I argue that the primary cause of this limitation comes from its psychological inaccuracy and failure to integrate the social embeddedness of economic agents. Economic models' inability to explain why agents are making choices limits their ability to predict future action. Undoubtedly it is important to make assumptions in any model that reduces the complexity of the world in order to make prediction easier and explanation simpler. However, the cost of these assumptions must not be so great so as to significantly damage predictiveness or limit explanatory power. The essential elements that rational choice theory misses are found in an agent's web of complex beliefs, desires, and commitments which extend beyond their own personal payoffs. An agent's desires are not merely formed within themselves but are developed within a society which enables them to build their own desires in relation to other agent's desires and needs. We must recognize these social influences if we are going to better explain and predict choices.

As I mentioned at the beginning, this view of agency was one that was held by Adam Smith, and his unique connection to both economics and philosophy offers the perfect opportunity to bridge economics paradigm of agency with a more psychologically accurate and social understanding of how agents really act. Taking a true Smithian stance towards economics can be the path to building a paradigm of agency within economics that acknowledges rational choice theory's strengths and weakness while taking a more holistic stance towards decision making. The intricacies of human agency are exactly what makes our decision making so unique, rather than being purely animalistic and desire driven. We have values, commitments and

responsibilities that obligate us to act beyond our preferences for items and goods. The way forward for economics starts with rectifying the understanding of Adam Smith's philosophy and the view he had for economics as a discipline.

The Adam Smith of today's economics is seen as a champion of free markets guided by self-interest, with groups like the Adam Smith Institute using his name as the mascot of free trade and market liberalization. I highlighted earlier on that there is an 'Adam Smith problem' when it comes to comparing this interpretation that many modern economists reading of *The Wealth of Nations* (WN) versus the writings found in *The Theory of Moral Sentiments* (TMS). The passage that talks about the invisible hand in WN states that human agents are being motivated by their own self-interest, but markets guide them to an outcome that is beneficial for society as well. Many introductory economic textbooks use the invisible hand passage as evidence that Smith believed in the good of self-interest and that he also was opposed to government intervention (Mankiw 2018). There is certainly some truth to this statement, but as has become a common trend in our discussion of economic theory, this view leaves lots on the table in terms of the depth of Adam Smith's contribution to the social sciences.

Primarily, in terms of philosophical considerations, the economic view portrays Smith as valuing efficiency as one of the greatest goods in an economy. As we will see, this fundamentally misses the point that Smith was trying to make by saying that everyone is guided by self-interest. Smith's idea of what constitutes self-interest is much more fleshed out in TMS and he emphasizes that there is a genuine consideration for other people built into human nature. He recognizes that every individual must take care of themselves and their family, as this is merely prudent. However, as the first line of TMS says, "How selfish soever man may be

supposed, there are evidently some principles in his nature, which interest him in the fortune of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it" (TMS I.i.1.1). Smith saw that there were considerations for people that were beyond merely their own self-interested payoff. He argued that no matter how selfish we may think people are, there is something fundamental about our nature that genuinely cares about and empathizes with the happiness of others, despite the fact that we get nothing out of it. He had a much broader definition of self-interest than just selfishness, and saw it as connected to a drive for survival rather than a drive to disregard others. Adam Smith's views of human agency and morality within TMS revolve around his conception of what he calls the 'impartial spectator' who guides us to evaluate our own actions through the lens of an impartial perspective.

Before explaining the impartial spectator and how it relates to our discussion of agency in economics though, let's look at the way that Smith describes human nature in TMS and how that can change our reading of passages within WN. He believed that all humans desire to be loved and praised by others. Smith argued that we all genuinely do care about what others think about us and are concerned about our desires being approved by others (TMS III.II.1). Importantly, he saw that there were influences beyond just our own self-interest that constrained our desires within our communities and society. Society is made up of lots of individuals who all approve or disapprove of other people's actions, which in turn helps individuals to change their behavior to better acknowledge the self-interest of others. Although Smith was writing this in the mid-1700s, he grasped an essential component of agency that continues to influence decision-making today. I would argue that the empirical evidence that we all have access to today showcases how individuals do genuinely care about other people and will change their behavior when put in differing social settings.

Smith's lengthy discussion of the ways that we approve and disapprove of people's actions in TMS, creates a very different perspective when reading WN. By thinking about Smith's emphasis in TMS on caring about others and being influenced by them, we can better assess what he meant in WN when he talked about self-interest. To avoid setting up a strawman, we will look at a more complex passage of Adam Smith rather than the singular (yes only once) time he utilizes the phrase 'invisible hand' in WN. One of the most famous passages in all of WN is about an interaction with a butcher, brewer or baker and it serves as a great example to rectify the 'Adam Smith problem':

"It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest. We address ourselves, not to their humanity, but to their self-love, and never talk to them of our own necessities but of their advantages. Nobody but a beggar chuses to depend chiefly upon the benevolence of his fellow citizens" (WN I.ii.26.7)

This passage can easily be interpreted as evidence for Smith claiming that people being self-interested is what motivates all economic outcomes – which leads to a view of agency similar to that of rational choice theory. Rational choice theory's paradigm emphasizes people being motivated by their own desires which fits nicely into the market system to explain how the market will guide all our self-interest toward the interest of society. Many economists read this passage as explaining how we never focus on benevolence of others to get what we want, but our own self-interest. We go to the butcher to get a cut of meat for dinner because we need to feed our family, and the butcher needs an income to feed his family so we can reach an agreement. We don't have to care about the butcher's interest outside of the way that it gives us the opportunity to buy dinner from him. If this passage is read as just describing how each agent is self-interested and therefore is competing to get what they want, then we reach a rational choice theory paradigm. The butcher has his own utility which he is trying to maximize, and the

customer has his own utility, which he is also trying to maximize. But that's not the point that Smith is making in this passage, and it's not what he believed guided agents.

Samuel Fleischacker, an Adam Smith scholar, explains how Smith is not talking about people appealing to their own desires, but appealing to *the desires of others*. The passage is actually talking about *other-oriented* concerns, rather than the individual's own interest. To use Fleischacker's words, "The main character, the character with whom we are supposed to identify, is the one who merely *appeals* to self-love… It is not at all clear that *this* character is self-interested" (Fleischacker 2004). The specific word choice is key for Smith - to say that 'we address ourselves not to *their* humanity, but to *their* self-love is to talk about empathizing with another person, not acting self-interestedly. This reading by Fleischacker is much more in tune with the philosophy that Smith wrote about in TMS. He heavily emphasized the role that empathy for others plays in our decision-making process, especially when it comes to making moral evaluations.

Understanding how Smith was utilizing self-interest in this context is critical to building his conception of agency. Smith is saying that it would be improper to expect that the butcher would just give you a cut of pork because you asked nicely. He must think about his own needs and feed his family - he has his own self-interest. On the other hand, it would be ridiculous for the butcher to expect to receive payment and then not give the customer a cut of meat. Both parties in an economic transaction must appeal to the *other person's* self-interest, not their own. This means that Smith is arguing for a process of decision making that takes not only your own desires into account, but also takes consideration for what other people expect and desire. When you read this passage in light of the writings in TMS, it becomes clear that he is not talking only about each individual's self-interest, but about how they recognize other people's self-interest.

Smith called this fellow feeling which eventually amounted to sympathy (but we would today call it empathy). We empathize with others when we are in social contexts and that drives us to think about concerns that are broader than merely our own personal payoffs.

Remember, Smith saw agency as being socially constituted and formed within a certain social environment. Humanity's desire to be loved drives us to seek approval from others for our actions, desires, and beliefs about the world. This approval process is a two-way street though, individuals are being evaluated by others and simultaneously evaluating others. Smith is talking about mutual empathic perspective taking in which, in an interaction between two people, each takes the others perspective and tries to bring it home to themselves. This is the process by which humans reach approval or disapproval of actions, by trying to understand them from their own perspective.

In addition to a desire to be loved and receive praise, Smith also identifies a love for praiseworthiness within all humans (TMS III.ii.2). This love of praiseworthiness does not come from praise though. Rather, Smith argues that "We must at least believe ourselves to be admirable for what [others] are admirable" (TMS III.ii.3). People seek to make themselves admirable in the ways that they think that other people are admirable. Their desires, goals, and choices are critically influenced by others because we grow up and learn in a society. In striving to be someone who is worthy of praise, one must engage in empathic perspective taking to view their actions from a perspective beyond their own.

Smith contends that one must be able to become an impartial spectator of their own behavior. In doing so, we reach a perspective that is neither ours, nor the other, but a spectator who is not partial to any party involved in the event (TMS III.ii.2). The power of sympathy in this context is to rise above one's own self-interest and see how your actions may affect another

person. We care about who we are as people, we have values, and we are committed to other people. The perspective of the impartial spectator is what helps us to fulfill those goals that we have for our conduct. Smith argues that we must recognize that we are "but one of the multitude, in no respect better than any other in it" (TMS III.3). Considering the impartial spectator, we change our behavior in order to be looked upon favorably from this perspective. This is arguably what we saw in all of those empirical studies which showcased individuals acting against what expected utility theory says was in their best interest but may not have been seen as fair.

The importance of this philosophy for our purposes in this paper, is to see that by adopting Smith's paradigm of agency into economics, we can work to better account for these influences that are not captured in the traditional rational choice model whilst still maintaining the essential elements that Smith brings to modern economics. Smith was a fierce critic of Thomas Hobbes for many reasons which included the fact that Hobbes came to his conclusion about human nature using a hypothetical situation *without society*. Hobbes' argument that human beings are driven only by desires and aversions left out the fact that people do live in society and are regularly influenced by other people. Smith argued that Hobbes misses the essential harmonious elements of society, which bring about virtuous feelings such as sympathy and commitment to each other (TMS VII.iii.2). This criticism that Smith had almost 250 years ago about the Hobbesian paradigm of agency, ironically still rings true for rational choice theory in economics.

The connection between economics and philosophy has been severed, but the application of economics still crucially depends upon some of conception of agency - and it's important to bring that into the foreground when analyzing why predictions are inaccurate. Mathematical functions give us impressive precision in economics but struggle predictively and have weak

explanatory power. Considering economics in the way that Smith envisioned means reinterpreting the way that we view agency, which directly impacts the outside constraints that we consider on decision-making. Smithian economics means assuming that we are all agents that have desires and beliefs and therefore we must acknowledge other people's agency. This starts from reading Smith as he was intended to be understood, and not placing economics in the Hobbesian picture of agency that sees human agents as merely motivated by their desires. Smithian economics is about economic models grounded in more accurate conceptions of agency and society. Accuracy with prediction dramatically increases when you work to incorporate these considerations such as we saw Cristina Bicchieri attempting to do. am not saying that economists as scientists should have anything to say about how the world ought to be, they should be looking at it from an objective standard, but they must be acutely aware that the agents whom they study have very much to say about the way the world ought to be.

Concluding Remarks

Alright, rational choice theory gets some predictions wrong, why should you care about the way that economics teaches their paradigm of agency? Well, it's been long recognized within economic research that you must control for the major of the undergraduates that you are studying[16]. Research shows that economics students play games in the laboratory differently, they play games more selfishly and more in line with the predictions of rational choice theory. John Carter, who was a Holy Cross professor of economics, sought to understand why this was occurring. He conducted research with freshmen who had only taken a couple months of macroeconomics as a control group, a collection of students from all other majors, and then senior economics students who had taken most of the undergraduate course load. He saw concretely that economics majors played ultimatum games more in line with rational choice theory predictions (although not exactly). They were willing to accept less from the proposer than the average student and gave less on average to a responder (Carter & Irons 1991). His most interesting finding though was that the freshmen econ majors who had only taken macro-economics were also behaving differently, even though neoclassical teachings of rational choice are primarily found in microeconomics.

So, his research concluded that economics students were different (more selfish), but he wasn't sure why. He thought more selfish students might be self-selecting into economics, which would simply say that economics is a more attractive major for selfish students. This research was done back in 1991, and Carter concluded the paper by suggesting that more research needed to be done on the topic because he wasn't convinced that it was merely self-selection. Since then, further research revealed that giving only a short lesson on economics dramatically increased

[16] Schmidt & Bauer 2008 is an example of such a control function in an econometrics paper.

selfish play in a variety of games (Ifcher & Zaraghamee 2018). Ifcher and Zaraghamee concluded that "We show that the lesson reduces efficiency and increases inequity in the [Ultimatum Game]. The results demonstrate that even a brief exposure to commonplace neoclassical assumptions measurably moves behavior towards self-interest" (Ifcher & Zaraghamee 2018). Their research, among others that came before them, gives more evidence to the claim that it is a taught behavior, not a self-selection phenomenon.

So, economics teaching a paradigm of agency that is focused on self-interest, despite the fact that it is possible to change the model to better account for altruism, is quite literally making students more selfish. That is why you should care about the paradigm of agency in economics.

I'm not saying that economics is only focused on self-interest, I showed you how they can potentially account for altruistic preferences. But it's being taught as if all agents are self-interested. The ways that we understand, describe, and evaluate our paradigms of agency is centrally important in order to pursue a society that strives to uphold our commitments to each other and cultivate students who care about others. The reason that this is so imperative for economics as a discipline is because economists regularly serve to advise politicians in the highest political offices in the United States. They make huge decisions that affect our daily lives every single day. This is why you should care about the paradigm of agency that exists in economics and the models they utilize that are fundamentally based upon it.

Bibliography

Alexandrova, Anna, and Robert Northcott. 2009. “Progress in Economics: Lessons from the Spectrum Auctions.” In The Oxford Handbook of Philosophy of Economics, edited by Don Ross and Harold Kincaid, 1st ed., 306–36. Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0011.

Ariely, Dan, and Jonathan Levav. 2000. “Sequential Choice in Group Settings: Taking the Road Less Traveled and Less Enjoyed.” Journal of Consumer Research 27 (3): 279–90. doi:10.1086/317585.

Batson, C. D. (2010). Empathy-induced altruistic motivation. In M. Mikulincer & P. R. Shaver (Eds.), *Prosocial motives, emotions, and behavior: The better angels of our nature* (pp. 15–34). American Psychological Association. https://doi.org/10.1037/12061-001

Bauer, Thomas K., and Christoph M. Schmidt. 2008. “WTP vs. WTA: Christmas Presents and the Endowment Effect.” SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1316403.

Binmore, K. G. 2007. Does Game Theory Work? The Bargaining Challenge. Economic Learning and Social Evolution 7. Cambridge, Mass: MIT Press.

Bernanke, Ben S. Non-Monetary Effects of the Financial Crisis in the Propagation of the Great Depression. National Bureau of Economic Research, 1983. National Bureau of Economic Research, https://doi.org/10.3386/w1054.

Berry, Christopher J., Maria Pia Paganelli, and Craig Smith, eds. 2013. *The Oxford Handbook of Adam Smith*. First edition. Oxford Handbooks. Oxford: Oxford University Press.

Bicchieri, Cristina, Ryan Muldoon, and Alessandro Sontuoso. 2018. “Social Norms.” In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/social-norms/.

Bicchieri, Cristina, and Alex Chavez. 2010. “Behaving as Expected: Public Information and Fairness Norms.” Journal of Behavioral Decision Making 23 (2): 161–78. https://doi.org/10.1002/bdm.648.

Carter, John R, and Michael D Irons. 1991. “Are Economists Different, and If So, Why?” Journal of Economic Perspectives 5 (2): 171–77. https://doi.org/10.1257/jep.5.2.171.

Cox, James C. 2004. “How to Identify Trust and Reciprocity.” Games and Economic Behavior 46 (2): 260–81. doi:10.1016/S0899-8256(03)00119-2.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. The Roundtable Series in Behavioral Economics. New York, N.Y. : Princeton, N.J: Russell Sage Foundation ; Princeton University Press.

Camerer, Colin F., and Ernst Fehr. 2002. “Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists.” SSRN Electronic Journal. doi:10.2139/ssrn.299143.

Cramton, Peter. 1998. “The Efficiency of the Fcc Spectrum Auctions.” The Journal of Law & Economics 41 (S2): 727–36. https://doi.org/10.1086/467410.

Davidson, Donald. 2004. *Problems of Rationality*. Oxford : New York: Clarendon Press ; Oxford University Press.

Fehr, E., and K. M. Schmidt. 1999. “A Theory of Fairness, Competition, and Cooperation.” The Quarterly Journal of Economics 114 (3): 817–68. https://doi.org/10.1162/003355399556151.

Fehr, Ernst, and Simon Gächter. 2002. “Altruistic Punishment in Humans.” Nature 415 (6868): 137–40. https://doi.org/10.1038/415137a.

Frankfurt, Harry G. 1971. “Freedom of the Will and the Concept of a Person.” The Journal of Philosophy 68 (1): 5. doi:10.2307/2024717.

Frankfurt, Harry G. 1988. The Importance of What We Care About. Cambridge [England] ; New York: Cambridge University Press.

Friedman, Milton. (1966) 2007. “The Methodology of Positive Economics.” In *The Philosophy of Economics: An Anthology*, edited by Daniel M. Hausman, 3rd ed., 145–178. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511819025.010.

Green, S.L. 2002. “Rational Choice Theory: An Overview.” Baylor University.

Guala, Francesco and Journal of Philosophy, Inc. 2006. “Has Game Theory Been Refuted?:” Journal of Philosophy 103 (5): 239–63. https://doi.org/10.5840/jphil2006103532.

Guala, Francesco. 2009. “Methodological Issues in Experimental Design and Interpretation.” In The Oxford Handbook of Philosophy of Economics, edited by Don Ross and Harold Kincaid, 1st ed., 280–305. Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0010.

Harrington, Joseph Emmett. 2015. Games, Strategies, and Decision Making. Second edition. New York, NY: Worth Publishers, A Macmillan Education Company.

Heath, Eugene. 2013. Adam Smith and Self-Interest. Oxford University Press. doi:10.1093/oxfordhb/9780199605064.013.0013.

Herfeld, Catherine. 2018. "From Theories of Human Behavior to Rules of Rational Choice." History of Political Economy50 (1): 1–48. doi:10.1215/00182702-4334997.

Herrnstein, R.j. 1990. "Rational Choice Theory: Necessary but Not Sufficient." *American Psychologist* 45 (3): 356–367. doi:10.1037/0003-066X.45.3.356.

Hobbes, Thomas, and Crawford Brough Macpherson. (1651) 1985. Leviathan. Penguin Classics. London: Penguin books.

Hume, David, and P. F. Millican. (1748) 2007. An Enquiry Concerning Human Understanding. Oxford World's Classics. Oxford ; New York: Oxford University Press.

Ifcher, John, and Homa Zarghamee. 2018. "The Rapid Evolution of Homo Economicus: Brief Exposure to Neoclassical Assumptions Increases Self-Interested Behavior." Journal of Behavioral and Experimental Economics 75 (August): 55–65. https://doi.org/10.1016/j.socec.2018.04.012.

Levin, Jonathan and Milgrom, Paul. 2004. "Introduction to Choice Theory." Stanford. https://web.stanford.edu/~jdlevin/Econ20202/Choice20Theory.pdf.

Jones, Martin K. 2021. "The Concept of Rationality in Introductory Economics Textbooks." *Citizenship, Social and Economics Education* 20 (1): 37–47. doi:10.1177/2047173421994333.

Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." Econometrica 47 (2): 263–91. doi:10.2307/1914185.

Kahneman, Daniel. 2013. Thinking, Fast and Slow. 1st pbk. ed. New York: Farrar, Straus and Giroux.

Mankiw, N. Gregory. 2018. Essentials of Economics. 8th edition. Australia ; Boston, MA: CENGAGE Learning.

McAfee, R. Preston, and John McMillan. "Analyzing the Airwaves Auction." Journal of Economic Perspectives, vol. 10, no. 1, Feb. 1996, pp. 159–75. DOI.org (Crossref), https://doi.org/10.1257/jep.10.1.159.

Mises, Ludwig Von. 1949. “The Prerequisites of Human Action.” In *Human Action*. The Mises Institute. https://mises.org/library/human-action-0/html/pp/614.

Morris, William Edward, and Charlotte R. Brown. 2022. “David Hume.” In The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2022. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2022/entries/hume/.

O’Neill, Onora. 2000. *Bounds of Justice*. Cambridge, U.K. ; New York: Cambridge University Press.

Paganelli, Maria Pia. 2013. Commercial Relations: From Adam Smith to Field Experiments. Oxford University Press. doi:10.1093/oxfordhb/9780199605064.013.0017.

Rabin, Matthew. 1993. “Incorporating Fairness into Game Theory and Economics.” The American Economic Review 83 (5): 1281–1302. https://www.jstor.org/stable/2117561.

Roth, A. E., T. Sonmez, and M. U. Unver. 2004. “Kidney Exchange.” The Quarterly Journal of Economics 119 (2): 457–88. https://doi.org/10.1162/0033553041382157.

Roth, A.E., T. Sönmez, M.U. Ünver, F.L. Delmonico, and S.L. Saidman. 2006. “Utilizing List Exchange and Nondirected Donation through ‘Chain’ Paired Kidney Donations.” American Journal of Transplantation 6 (11): 2694–2705. https://doi.org/10.1111/j.1600-6143.2006.01515.x.

Reiss, Julian. 2013. *Philosophy of Economics: A Contemporary Introduction*. Routledge Contemporary Introductions to Philosophy. New York, NY: Routledge.

Rosenberg, Alex. 2009. “If Economics Is a Science, What Kind of a Science Is It?” In The Oxford Handbook of Philosophy of Economics, edited by Don Ross and Harold Kincaid, 1st ed., 55–67. Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0003.

Satz, Debra, John Ferejohn, and Journal of Philosophy Inc. 1994. “Rational Choice and Social Theory:” Edited by John Smylie. Journal of Philosophy 91 (2): 71–87. doi:10.2307/2940928.

Sen, Amartya. 2013. The Contemporary Relevance of Adam Smith. Oxford University Press. doi:10.1093/oxfordhb/9780199605064.013.0029.

Sen, Amartya K. 1977. “Rational Fools: A Critique of the Behavioral Foundations of Economic Theory.” *Philosophy & Public Affairs* 6, no. 4, 317–44. http://www.jstor.org/stable/2264946.

Simon, Herbert A. 1955. “A Behavioral Model of Rational Choice.” *The Quarterly Journal of Economics* 69 (1): 99. doi:10.2307/1884852.

Simon, Herbert A. 1990. “Bounded Rationality.” In *Utility and Probability*, edited by John Eatwell, Murray Milgate, and Peter Newman, 15–18. The New Palgrave. London: Palgrave Macmillan UK. doi:10.1007/978-1-349-20568-4_5.

Smith, Adam, D. D. Raphael, A. L. Macfie, and Adam Smith. 1982. The Theory of Moral Sentiments. The Glasgow Edition of the Works and Correspondence of Adam Smith 1. Indianapolis: Liberty Classics.

Stein, Edward. 1996. *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science*. Clarendon Library of Logic and Philosophy. Oxford : Oxford ; New York: Clarendon Press ; Oxford University Press.

Stueber, Karsten R. 2006. *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. Cambridge, Mass: MIT Press.

Taylor, Charles. 1976. “Responsibility For Self.” In *The Identities of Persons*, edited by Amélie Oksenberg Rorty, 281–299. Topics in Philosophy 3. Berkeley, Calif.: Univ. of California Press.

Thaler, Richard. 1980. “Toward a Positive Theory of Consumer Choice.” Journal of Economic Behavior & Organization 1 (1): 39–60. https://doi.org/10.1016/0167-2681(80)90051-7.

Thaler, Richard H. 2015. Misbehaving: The Making of Behavioral Economics. First edition. New York: W.W. Norton & Company.

Thaler, Richard H., and Cass R. Sunstein. 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. New Haven: Yale University Press.

Tversky, Amos. 1969. “Intransitivity of Preferences.” Psychological Review 76 (1): 31–48. https://doi.org/10.1037/h0026750.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. “The Causes of Preference Reversal.” The American Economic Review 80 (1): 204–17. https://www.jstor.org/stable/2006743.

Von Neumann, John, and Oskar Morgenstern. 1944. The Theory of Games and Economic Behavior, 2nd ed. Princeton, NJ: Princeton University Press, 1947.

Wheeler, Gregory. 2020. “Bounded Rationality.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/bounded-rationality/.

Wight, Jonathan. 2002. *Saving Adam Smith: A Tale of Wealth, Transformation, and Virtue*. Upper Saddle River, NJ: Prentice Hall

Woodward, James F. 2009. "Experimental Investigations of Social Preferences." In The Oxford Handbook of Philosophy of Economics, edited by Don Ross and Harold Kincaid, 1st ed., 189–222. Oxford University Press. doi:10.1093/oxfordhb/9780195189254.003.0007.