By: Garrett Pendergraft

GAME THEORY

Abstract: Game theory involves deliberating about what to do in light of what other people are likely to do. One of the central frameworks of game theory is the prisoner's dilemma, in which participants who make rational choices end up in sub-optimal outcomes. Using the prisoner's dilemma to model competition between firms sets the stage for a new and promising approach to business ethics: the market failures approach.

SECTION 1: THEORY

What is Game Theory?

Game theory is an approach to solving problems that attempts to model real-world interactions using artificial frameworks. These artificial frameworks can be described more precisely than the situations they are trying to represent, which facilitates more fruitful theorizing. This theorizing focuses on **economic agents** who operate according to certain **preferences** and produce certain **outcomes** (Ross 2021). Game-theoretic approaches are successful to the extent that they provide insight into the real-world situations that they are modeling.

What makes game theory distinctive from other ways of modeling reality is that it focuses on the reasoning of economic agents in light of the reasoning of *other* economic agents. Ross (2021) offers an illustrative example of the complexities that arise when our reasoning has to take into account the reasoning of other agents. Imagine that you are trying to decide the best way to cross a river. The closer bridge is more dangerous; the safer bridge is farther away. Deciding which bridge to cross simply requires calculating the risks and weighing them against your preferences. If you are risk-averse, and don't mind a longer journey, then it's worth taking the extra time to cross the safer bridge. But if you're in a hurry, then the risk might be worth it. But now suppose that someone on the other side is trying to intercept you. In this new situation you have to take into account not only your own preferences, but also the deliberations of your adversary. What will they expect you to do, and how can you adjust your behavior so as to violate those expectations? This becomes even more complex when you realize that they already know that you're going to try to violate their expectations, and will adjust their own reasoning process accordingly. Thus, introducing the other agent into the bridge situation creates a dilemma.

One of the central frameworks—if not *the* central framework—in game theory is that of the *prisoner's dilemma*. This framework can be used to represent various choice situations and explain why agents in those situations tend to make choices that lead to suboptimal outcomes.

Suppose that you and a stranger are selected at random to play a high-stakes coordination game in which you are both given the opportunity to take home a share of $1,000,000. The catch is that you each have to decide, individually, whether you are going to *split* or *steal*. If you choose split, then you are choosing to **cooperate**: you are offering to split the money, so that

each of you gets $500,000. If you choose steal, then you are choosing *not* to cooperate (we might call this a choice to **defect**): you are signaling a desire to take the entire $1,000,000 for yourself. Since there are two of you in this game, and each of you has two choices, there are four possible outcomes:

	Split	*Steal*
Split	$500,000 / $500,000	$0 / $1,000,000
Steal	$1,000,000 / $0	$0 / $0

If you both choose split, then you both take away $500,000. If you offer to split but the other player chooses steal, then they take all the money and you get nothing. If, on the other hand, the other player offers to split but you choose to steal, then *you* take all the money while the other player gets nothing. And if you both choose to steal, then you both get nothing.

The essential feature of this game is that a rational choice leads to a sub-optimal outcome. Here's how you both might reason about the game:

> If the other player chooses to split, then I can make more money ($1,000,000 instead of $500,000) by choosing to steal; so if the other player splits then it would be better for me to steal. But if the other player chooses to steal, then there's no way I can take home any money; and I don't want to reward them for making the selfish choice! So if the other player steals then it would be better for me to steal too. In short, no matter what the other player does, it would be better for me to choose steal.

Unfortunately, if you both reason in this way, then you both end up with nothing. The best overall outcome is for you both to end up with $500,000, but thinking rationally about the game pushes you both toward the sub-optimal outcome.

This is an example in which the outcomes are (potentially) positive, but a similar game can be generated with negative outcomes instead. (In fact, the reason why it's called a *prisoner's* dilemma is because it was originally conceived as a situation in which the outcomes are prison sentences rather than monetary rewards.)

We can describe the general structure of a prisoner's dilemma using the following generic payoff matrix (Kuhn 2019):

	Cooperate	*Defect*
Cooperate	R / R	S / T
Defect	T / S	P / P

In this payoff matrix, T stands for the *temptation* to defect and take all the benefit for yourself; R stands for the *reward* you'll receive if you both cooperate; and P stands for the *punishment* you'll receive if you both defect. S stands for *sucker*: the reason why most people choose to defect is that they don't want to be the sucker who cooperates while the other person is defecting. These payoff values sort out as follows: T > R > P > S. (It's also possible for P and S to be equal, as in the first example.) Thus, any choice situation in which the outcomes can be ranked according to the generic payoff matrix above counts as a prisoner's dilemma.

[insert animated explainer video]

Background

Although game theory wasn't formalized until the 1940s, hints of it show up in various episodes in military history (Ross 2021). For example, in the *Symposium*, Socrates describes a dilemma in which soldiers realize that, win or lose, the better thing for them would be to desert the army. And Hernán Cortés was probably aware, at least implicitly, of the possibility of this type of reasoning when he destroyed his ships to prevent his outnumbered Spaniards from deserting (Hassig 2006, 77). Destroying the ships restructures the situation so that desertion is no longer the rational option.

Once game theory became an active area of research in the 1940s, discussion of the prisoner's dilemma arose relatively quickly. It was introduced by John Nash in his dissertation, but the label comes from Nash's advisor, Albert W. Tucker, who mentioned the thought experiment in a 1950 lecture at Stanford (Peterson 2015, 1). (Nash, who's featured in the biopic *A Beautiful Mind*, went on to win a Nobel Prize in Economics.) Discussion of the prisoner's dilemma gained steam in the 60s and 70s, and a Google Scholar search for "prisoner's dilemma" in 2022 returns almost 60,000 results.

Criticisms and Limitations

A model, at least as we're understanding the term, is a representation of reality; and every representation is necessarily incomplete (or at least imperfect). The prisoner's dilemma is no different, so it will be more or less open to criticism depending on how much insight it's able to provide into a particular situation. It's supposed to help us explain and predict a wide variety of phenomena, but some philosophers (e.g., Northcott and Alexandrova (2015)) have argued that these explanatory benefits are minimal.

This is probably a minority view, however; most theorists see a continued focus on the

prisoner's dilemma as fruitful and worthwhile. One of the primary disputes among those who take the prisoner's dilemma seriously is whether a decision to cooperate can be rationally justified after all. This often takes the discussion into different variations on the dilemma, such as those involving more than two players or more than one iteration. See Gauthier (2015) for an argument that cooperation can be rational even in a one-off prisoner's dilemma; although, as Peterson (2015, 10) notes, Gauthier's argument does rely on an atypical understanding of **practical rationality**.

New and Emerging Directions

One context in which the prisoner's dilemma shows up, perhaps counterintuitively, is that of competitive sports (Heath 2007). Consider a group of sprinters who are trying to decide how much to train. If none of them train, then the person who is naturally the fastest will win. But if someone who isn't naturally the fastest trains, then she can win in virtue of her training. The other runners will realize this, and will probably decide to train as well. But once they're all training, then, assuming a roughly equal training regimen, the person who is naturally the fastest will probably win. So, in the end, everyone has put a significant amount of time, energy, and money toward training, but the result is basically the same as it would have been had they all simply agreed not to train! Of course, in the case of elite athletics, this outcome is sub-optimal for the participants but produces a great deal of net benefit to society: the fans get to enjoy a much higher quality product in virtue of all the participants making the choice to defect in the prisoner's dilemma that they face.

Another area in which insights from game theory have started to emerge is that of *business ethics*. According to conventional wisdom, "business ethics" is a contradiction in terms: to engage in business is to pursue one's self-interest, but acting ethically involves prioritizing the interests of others. How could the two possibly be reconciled? Fortunately, as we will see in Section 2, constructing a viable framework for business ethics does not require attempting to reconcile self-interest with altruism. If we represent competition between firms as a prisoner's dilemma, we can see how some of the most important ethical principles in business arise directly from the nature of business itself.

Knowledge and Understanding Check

Question 1: What is distinctive about game theory as a field of inquiry?

Answer choices	*Correct?*	*Feedback*
Game theory applies the probability calculus to decision making.	Incorrect	Game theory does often apply the probability calculus, but that's not what makes it distinctive.
Game theory focuses on reasoning that is done in light of the reasoning of other agents.	Correct	Game theoretic reasoning always includes considerations about what another agent will do, and how they will reason about that choice.
Game theory hasn't had much impact outside the field of economics.	Incorrect	Game theory has had a huge impact in economics, but also in philosophy, law, political science, sociology, anthropology, biology, and more.
Game theory experienced a golden age in the 50s, but was then eclipsed by other methods and frameworks.	Incorrect	Interest in game theory has continued to grow since it became an object of academic inquiry in the 1950s.

Question 2: What is it that makes the prisoner's dilemma a dilemma?

Answer choices	*Correct?*	*Feedback*
A rational choice at the group level leads to an overall worse outcome for one of the individuals.	Incorrect	The relationship actually goes in the opposite direction: a rational choice at the *individual* level leads to an overall worse outcome for the group.
There's no best outcome for the group as a whole.	Incorrect	There is an overall best outcome for the group, it's just that individual rational choices will prevent that best outcome from occurring.
A rational choice at the individual level leads to an overall worse outcome for the group.	Correct	What appears to be acceptable self-interested reasoning by an individual inside of a prisoner's dilemma leads to an outcome that is worse for everyone.
There's no possibility of rational choice at the individual level.	Incorrect	It's possible to make a rational choice, but that rational choice leads to a sub-optimal outcome.

Question 3: How should we evaluate the usefulness of the prisoner's dilemma as a theoretical framework?

Answer choices	*Correct?*	*Feedback*
By how closely it hews to John Nash's original formulation.	Incorrect	Theoretical developments since Nash's introduction of the dilemma have helped it deliver additional insights.
By how many mentions it gets in Google Scholar.	Incorrect	Scholar mentions are a useful proxy for popularity, but not necessarily usefulness.
By how well it helps us decide what to do in the bridge example.	Incorrect	The prisoner's dilemma is designed to provide insight in a wide range of situations.
By how much insight its representation of reality is able to provide.	Correct	The hope for the prisoner's dilemma is that it can model actual situations in a way that generates novel insights about those and other situations.

Key Terms

cooperate [in a coordination game, to cooperate is to make the choice that the other individual would want you to make]

defect [in a coordination game, to defect is to make a choice that the other individual would not want you to make]

economic agent [an agent who is defined by a set of preferences that determine how they will behave in a given situation]

practical rationality [the set of norms and practices that govern our practical reason—i.e., our reasoning about what to do]

preference [a measure of the utility that an economic agent assigns to an outcome, which in turn can be used to explain and predict the behavior of the agent]

outcome [a state of affairs that results, wholly or in part, from a decision made by an economic agent]

utility [the value that an economic agent attributes to an outcome]

References

Gauthier, D. (2015). How I learned to stop worrying and love the prisoner's dilemma. In M.

Peterson (Ed.), *The prisoner's dilemma* (pp. 35–53). Cambridge University Press.

Hassig, R. (2006). *Mexico and the Spanish conquest* (2nd ed). University of Oklahoma Press.

Heath, J. (2007). An Adversarial Ethic for Business: Or When Sun-Tzu Met the Stakeholder. *Journal of Business Ethics*, *72*, 359–374.

Kuhn, S. (2019). Prisoner's Dilemma. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2019). https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/

Northcott, R., & Alexandrova, A. (2015). Prisoner's dilemma doesn't explain much. In M. Peterson (Ed.), *The prisoner's dilemma* (pp. 64–84). Cambridge University Press.

Peterson, M. (2015). Introduction. In M. Peterson (Ed.), *The prisoner's dilemma* (pp. 1–15). Cambridge University Press.

Ross, D. (2021). Game Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). https://plato.stanford.edu/archives/fall2021/entries/game-theory/

SECTION 2: THEORY IN PRACTICE

When to Apply Game Theory vs. an Alternative

If you are deliberating about what to do, and in your deliberations you are taking into account what other agents might do as well, then you are engaging in game theory. So the primary question is not *when* to engage in game theory, but what kinds of simplifying assumptions to make in the numerous and varied contexts in which you need to deliberate in light of the deliberations of other agents.

Thus game theory is often relevant from the perspective of first-person deliberation. But it can also be relevant when trying to explain behavior, or social dynamics more broadly, from a third-person perspective. And as we will now see, it can even be relevant when trying to justify **normative constraints** on market transactions and related activities—i.e., it can even be relevant when engaging in business ethics.

An exciting new approach to business ethics, built in part on insights from game theory, emerged in the 21st century. This new approach—the **market failures approach**—was introduced and developed by Joseph Heath (2004, 2014).

How to Apply Game Theory: the Market Failures Approach

The starting point for Heath's theory is recognizing that a free market can be modeled as a prisoner's dilemma. Imagine two competing suppliers of some product—say, a ball bearing. If the two suppliers produce bearings of the same quality and sell them for the same price, then

they will both make a certain amount of profit. But if one of the suppliers decides to sell their bearing for slightly less, then they will increase their profit by taking market share from the other supplier. The other supplier, recognizing this, will probably lower their price as well. This will produce what's called a **race to the bottom**, where both suppliers will lower their prices to the lowest point at which they can sustain their business. They are, in effect, facing a prisoner's dilemma and both choosing to defect. As in the case of competitive sports (Heath 2007), this situation is bad for the suppliers but good for society in general: because the suppliers are making the rational choice to defect, consumers are getting their ball bearings for a lower price than they would have if the suppliers had cooperated with each other. In other words, the pricing mechanism in a free market operates by forcing sellers into a prisoner's dilemma. And this is actually a good thing: their involvement in a prisoner's dilemma is what enables the market to allocate resources efficiently. If sellers cooperated rather than defecting, then the market would fail to do what it's supposed to do. (This is one reason why price-fixing is illegal.)

This, again, is an oversimplified (and idealized) picture of how markets work. No actual market works in exactly this way, but it is how markets *should* work. The nature of a market is to distribute goods and services in the most efficient way, and anything that hinders that efficiency violates the purpose of the market. Anything that undermines the efficiency of the market does so by producing a **market failure**.

This insight about the function of the market provides the foundation for Heath's market failures approach to business ethics. On this approach, business ethics is not some attempt to reconcile self-centered behavior with altruism; instead, it starts with a recognition that participants in a market shouldn't be doing anything that tends to produce a market failure. This maxim—*It's unethical to perform an action that tends to produce a market failure*—is the fundamental principle of business ethics. Price-fixing is unethical because it produces market failure. False advertising is unethical because it produces market failure. (The price mechanism can't operate at maximum efficiency if consumers are given misleading information about whether and to what extent a product will satisfy their desires.) Excessive pollution is unethical because it produces a market failure. (Excessive pollution creates a **negative externality**, which occurs when a producer is paying less than the full production cost of the goods or services they're providing; and a market can't operate at maximum efficiency when producers aren't paying a full production cost.)

This core principle of the market failures approach implies various specific directives. Here are a few of them, which help constitute what Heath (2014, 37) dubs "the market failures code":

- Minimize negative externalities.
- Compete only through price and quality.
- Reduce information asymmetries between firm and customers.
- Do not exploit diffusion of ownership.
- Avoid erecting barriers to entry.

- Do not oppose regulation aimed at correcting market imperfections.

In essence, the market failures approach is making a conditional claim: *If* you are engaging in free market transactions, then certain constraints apply to you. These constraints are not imposed from some external moral system, but instead arise from the nature of the market itself.

So: one benefit of the market failures approach to business ethics is that it dissolves the alleged internal tension in the concept of business ethics. Another benefit is that it provides a better framework for thinking about **corporate social responsibility**.

For a long time, business ethicists who wanted to theorize about corporate social responsibility were forced to choose between the **shareholder theory** (Friedman 1970) and the **stakeholder theory** (Freeman 1979). According to the shareholder theory, managers of a firm have a fiduciary duty only to its owners, which means that there's basically no such thing as corporate social responsibility. According to the stakeholder theory, managers have fiduciary duties not just to owners, but also to employees, suppliers, and other stakeholders. Heath (2006) explains why both of these approaches are flawed and offers a superior alternative.

In short, a focus on game theory leads to a better framework for thinking about business ethics. Recognizing that free market transactions can be modeled as a prisoner's dilemma helps explain how genuine ethical constraints on the behavior of a firm arise organically from the nature of the market itself.

Field Report: Theory in Action

[insert interview with Mariam Thalos]

Practice with Interactive Scenarios

Scenario 1: Consider a situation involving a competition between wizards (Finkel 2018): You are a wizard (Wizard 1) who is facing off against two other wizards (Wizard 2 and Wizard 3) in a duel. Wizard 2's wand works 70% of the time and Wizard 3's wand works 90% of the time. Starting with you, each wizard will take turns casting an attack spell (or deciding to pass), in order, until only one wizard is left standing. You have a choice between three wands: one that works 60% of the time, one that works 80% of the time, or one that works 100% of the time. Which wand should you choose, and whom (if anyone) should you attack when it's your turn?

Answer choices	*Feedback*
100% wand	Your initial thought might be to choose the 100% wand—it is, after all, the most effective instrument. But if you choose this wand, then you will always be the biggest target because you have the most powerful wand. So, perhaps

	surprisingly, this wand is not the best choice.
80% wand	The problem with this wand is that it makes you a greater threat than Wizard 2 (the 70% wizard). Wizard 3 will target you since you're a greater threat than Wizard 2.
60% wand	Although it seems counterintuitive, this wand gives you the best chance of survival. If you choose this wand and then pass (i.e., refrain from attacking when it's your turn), then the other two wizards will target each other until one of them is defeated. (And then you can target the remaining wizard, knowing that you have a 60% chance of defeating them with your attack.)

Because your choice in this scenario has to be sensitive to the reasoning process of the other wizards, it turns out that the best choice is to make yourself the weakest opponent.

Scenario 2: You and a friend are scheduled to compete in the state final of the 100-meter race. (You two are clearly the top contestants, so you don't need to worry about the rest of the field.) The race takes place in a month. As it stands now, your friend is slightly faster than you; so if the race were to be held today instead, then they would probably just barely beat you. How much should you train over the next month?

Answer choices	*Feedback*
Don't train at all	If you don't train at all, then you'll have basically no chance of winning the race. Assuming that you want to win the race, this is not a good choice.
Train a moderate amount	If you train a moderate amount, then there's a slight chance that you'll win the race—if, for example, for some reason your friend decides not to train at all.
Train an extreme amount	Training an extreme amount gives you a much better chance of winning. Unless your friend matches your extreme training amount, you'll probably win. On the other hand, if your friend *does* match your training amount, then you will both have essentially trained for nothing: the outcome will be basically the same as it would have been if both of you had decided not to train at all.

Thus, to the extent that you both want to win, you will both be inclined to waste your time by training without making a difference to the ultimate outcome. However—the interesting feature of this scenario (which makes it similar to the way a free market works) is that the more you both "waste your time," the better things are for fans and others who are interested in the outcome. By

training, you are essentially losing a prisoner's dilemma for the benefit of society.

Scenario 3: You own a factory that emits a significant amount of air pollution. The local government has threatened to impose some costly regulations unless the average daily air pollution index stays below 4 (on a 10-point scale). After talking with the other factory owners, you all estimate that if each of you spent about $100,000 on pollution mitigation measures, you could be reasonably certain that the air quality would remain good enough to avoid the even costlier regulations. So you and the other owners agree to implement the mitigation measures. After making the agreement, however, you start to have second thoughts; you start to wonder if keeping the agreement is really in your best interests. What are your options here?

Answer choices	*Feedback*
Stick to the agreement	If you stick to the agreement, then you are doing your part to avoid the regulations. However, if the other factory owners don't cooperate, then it's likely that the air quality will deteriorate to the point where the government imposes the costly regulations. If this happens, then you will have spent money on the mitigation measures but you'll still have to face the costly regulations.
Violate the agreement	If you violate the agreement but the other factory owners stick to it, then you've gotten all of the benefits without any of the costs. However, if enough of the other factory owners reason in the same way, then the air will become polluted and costly regulations are sure to follow—and in that case you all would have been better off if you had just stuck to the agreement.

The prisoner's dilemma structure is clear in this case. It's not hard to see how you can reason yourself into violating the agreement, but if enough people do that then you'll all end up worse off overall. The question, then, is how to modify the incentive structure so that it's rational for everyone to stick to the agreement and thus end up in a better position overall.

<u>Discussion questions</u>

1. Can you think of a recent situation in which you were using game theory by deliberating in light of how the other individuals in the situation were deliberating? Were you happy with the choice that you made?
2. Based on what you know about the prisoner's dilemma, is it ever rational to cooperate? What would you need to change about the choice architecture in order to incentivize cooperation?

3. As we saw above, one explanation for the training practices of elite athletes is that they're stuck in a prisoner's dilemma. Why are so many athletes willing to stay in the prisoner's dilemma despite the fact that the majority of the benefit goes to society?
4. Take a look at the market failures code above. Can you think of any recent examples in which corporations have violated this code? Are there other ways in which bad behavior in business can be thought of as "unhealthy" competition, analogous to unhealthy competition in sports?

Key terms

corporate social responsibility [the responsibility, if any, that a corporation has to go beyond its purpose of making a profit and pursue other purposes that have a social benefit]

market failure [any situation that prevents market transactions from producing the most efficient allocation of goods and services]

market failures approach [an approach to corporate social responsibility (and business ethics more generally) that derives normative constraints from the role that the market plays in the efficient allocation of resources]

negative externality [a cost that arises from economic activity and is paid by a third party rather than by the party that generates the cost]

normative constraint [a principle about how things *ought* to be, in contrast to a descriptive claim merely about the way things *are*]

race to the bottom [a dilemma in which all participants make a rational choice that nevertheless leads to a worse outcome, where that worse outcome reinforces their incentive to make further choices that lead to successively worse outcomes]

shareholder theory [a theory of corporate social responsibility according to which a manager's only obligation is to make as much money as possible for the owners of the firm]

stakeholder theory [a theory of corporate social responsibility according to which a manager has obligations not only to the owners of the firm, but also to additional stakeholders such as employees, customers, and the broader community]

References

Finkel, D. (2018, May 22). *Can you solve the wizard standoff riddle?* TED-Ed. https://ed.ted.com/lessons/can-you-solve-the-wizard-standoff-riddle-daniel-finkel

Freeman, R. E. (1979). A stakeholder theory of the modern corporation. In T. L. Beauchamp & N. E. Bowie (Eds.), *Ethical theory and business*. Prentice Hall.

Friedman, M. (1970, September 13). The Social Responsibility of Business Is to Increase its Profits. *The New York Times Magazine*.

Heath, J. (2004). A Market Failures Approach to Business Ethics. In B. Hodgson (Ed.), *Studies in Economic Ethics and Philosophy* (Vol. 9). Springer.

———. (2006). Business Ethics without Stakeholders. *Business Ethics Quarterly*, 16, 533–57.

———. (2007). An Adversarial Ethic for Business: Or When Sun-Tzu Met the Stakeholder. *Journal of Business Ethics*, *72*, 359–374.

———. (2014). *Morality, Competition, and the Firm: The Market Failures Approach to Business Ethics*. Oxford University Press.