Game Theory and Decision Theory

Antony Eagle

Choices, Models and Morals » Lecture 4

Contents

Games and Strategic Interactions

Preferences, Actions, and Choices

- The model for rational choice we've been working with from lecture 2 is one where a player makes a **choice about an action to perform** against the background of some unknown state of nature.
- The values assigned to outcomes and the probabilities assigned to those states dictated the agent's **preferences among actions**.
- We saw that this model faced some difficulties in cases of **strategic interaction**, such as the 'Israeli War' example (Bar-Hillel and Margalit 1972).
  - A strategic interaction is one where the relevant states of nature are the **unknown choices of other rational agents**.
- We briefly investigated using **conditional probabilities** to model rational choice in such cases, but for many preference-first theorists in economics and elsewhere, this reification of probabilities and values is both **problematic** and **unnecessary**.

Game Theory

- It is unnecessary, they say, because there is a model of strategic interaction where all you need to know is the **preferences of the interacting players for outcomes** – no probabilities required.
  - It may be argued that the assignment of conditional probabilities in the Israeli War example **presupposes** such knowledge of preferences: the reason that war is **more likely** if Israel remains is because Egypt **prefers** war to peace in that case.
- **Game theory** is the theory of rational choice amongst actions when competing against other rational agents, and where what is rational for you might depend on their choices – in which case, the actions are called **strategies**.
- The idea is that we can generate normative principles, and descriptive predictions for rational agents, by considering the structure of preferences alone.
  - There is a potential here for a clash with individual decision theory: we return to this issue below.

The Structure of a Game

- A **game** has:
  1. **Players** who interact; these may be individuals or not, and there may be any number more than 2;
  2. **Strategies** which each players adopt for the whole game, though these can be **conditional**, varying in the choice depending on the actions of other players;
  3. **Outcomes** for each player, which emerge from the strategies each have chosen; in simple games, the strategies fully determine the outcomes.
  4. **Pay-offs** for each player in each outcome, that represent, perhaps indirectly, the **value** they get from that outcome being realized; these are dictated by the player's preferences over outcomes.
- In order for an interaction to be a game, it needs to be assumed that all the players **know** they are playing a game, and which game it is.
- There may be no rational strategy if you can't rely on other players playing the game. So we usually assume something stronger: **common knowledge of rationality**, namely, that every player knows the game they are playing, that every player knows this of every player; that every player also knows that every player knows this of every player; and so on *ad infinitum*.

A Simple Game

- Suppose two players are trying to decide which concert to go to. They make individual choices between Mozart and Mahler and each prefers Mozart to Mahler.
- Let's assume they make their choices **simultaneously**, so (in effect) they reveal their strategies to one another once the outcome is realized (perhaps they just turn up to their chosen concert and see if the other is present).
- This is not a 'sequential' game, and can be represented in strategic form (Reiss 2013: 57–59), as follows:

Table 1: Mozart or Mahler?

1↓ / **2**→	Mozart	Mahler
Mozart	(2, 2)	(1, 0)
Mahler	(0, 1)	(0, 0)

- Going to the Mozart concert is the best outcome for each – i.e., it is best for all if the collection of strategies {**1** goes to Mozart, **2** goes to Mozart} is realized.

Nash Equilibria; or, No Regrets

- One way to see that the {Mozart, Mozart} strategy is good for all jointly is to consider whether the players would have **regrets** once they see what the chosen strategies turned out to be.
- Suppose **1** went for Mahler, and **2** went for Mozart. Then **1** would have regret; given what **2** actually did, **1** would have preferred to go to Mozart. Suppose both go to Mahler; that's better for each than going to Mahler alone, but both would have regret.
- A set of strategies where no one would have regrets given that the other players played strategies in that set is a **Nash equilibrium** of the game (Reiss 2013: 57): that is, a set of strategies

> such that each player's … strategy maximises his pay off if the strategies of the others are held fixed. (Peterson 2017: 241; see also Reiss 2013: 58)

  - {Mozart, Mozart} is a Nash equilibrium; each player is happy to have played what they did, given what the others did.
  - This is the **unique** Nash equilibrium; so, we might say, it is the only way to avoid regret when you find out what strategy the other player adopted, and hence is the rationally mandatory act (Reiss 2013: 56).

Modified Mozart or Mahler

- A game with a Nash equilibrium need not have just one.
- Suppose we modify the pay-offs for player **1** and player **2** in Mozart or Mahler, so that now each would prefer to go together than separately:

Table 2: Modified Mozart or Mahler? (Osborne and Rubinstein 1994: 16)

1↓ / **2**→	Mozart	Mahler
Mozart	(2, 2)	(0, 0)
Mahler	(0, 0)	(1, 1)

- Here the solution {Mahler, Mahler} is also a Nash equilibrium: if player **2** plays Mahler, player **1** will prefer to have played Mahler; likewise, if player **1** plays Mahler, player **2** will prefer to have played Mahler too.
- So while both players will prefer to go together to the Mozart concert, the also prefer each other's company to Mozart alone.
- What should players do in this case?

Game Solutions and Nash Equilibria

- A **game solution** is a set of strategies that rational players **will** (or **would**) jointly play – because it is what they would do, there can be only one game solution.
- There is widespread agreement that **rational players will jointly converge on some Nash equilibrium**: no one could rationally have regrets.
  - So any game solution must be a Nash equilibrium.
- But as we just saw, there is no guarantee that there is only one Nash equilibrium, so if there is a game solution in such a game, then some Nash equilibrium is not a game solution.
  - {Mozart, Mozart} might be the unique solution in Modified Mozart or Mahler without being a unique Nash equilibrium.
- Maybe it is **Pareto optimality** that makes for a unique solution here (Reiss 2013: 68)?
  - An outcome X is a **Pareto improvement** over Y iff someone strictly prefers X to Y, and no one strictly prefers Y to X. An outcome is **Pareto optimal/efficient** iff there is no outcome that is a Pareto improvement over it.

Easy cases

- In the original Mozart or Mahler case, the unique Nash equilibrium is the best outcome for all, so individual and jointly rationality converge.
  - Going to the Mozart concern is the dominant strategy for each player, as well as being in the Nash equilibrium.
- And we can also imagine a case where there is no Nash equilibrium: no matter what, someone will regret their choice, and hence there is no jointly rational strategy (though individually rational choices might still exist).

Table 3: Love and Loathe

1↓ / **2**→	Café	Bar
Café	(2, 1)	(1, 2)
Bar	(1, 2)	(2, 1)

- In Love and Loathe, whatever player **2** does, player **1** wants to do; but whatever player **1** does, player **2** wants to the do the opposite.
  - This is a **strictly competitive** game, one 'where the interests of the two players are diametrically opposed' (Osborne and Rubinstein 1994: 17). Another example is

Cooperation and Rational Choice

Cooperation and Strategy

- Any moral theory faces a fundamental ethical question: **why be moral?**
- Game theory promises a partial answer: **it is sometimes rational to engage in pro-social and cooperative behaviour**.
  - Compared to individual decision theory, where everyone is invited to consider only their own **self-interest**, in game theoretic analyses the players are forced to consider each other as rational agents. There is something fundamentally **interpersonal** and other-regarding in game solutions that makes it apt to speak of morality in this context.
- Can cooperation rationally emerge from strategic interaction (Skyrms 1996)? In Rich Friend (below), the unique Nash equilibrium involves player 1 opting to go to a café which has a lower average payoff for them than going to a fine dining restaurant; in that sense, they sacrifice their self-interest to get a good joint outcome.

Table 4: Rich Friend

1↓ / **2**→	Café	Fine dining
Café	(4, 4)	(1, 0)
Fine dining	(3, 2)	(6, 1)

The Stag Hunt

- A small tweak on Mozart-or-Mahler gives us the Stag Hunt:

Table 5: Stag Hunt

1↓ / 2→	Hunt Stag	Hunt Hare
Hunt Stag	(5, 5)	(0, 2)
Hunt Hare	(2, 0)	(1, 1)

- This game gets its name from a passage in Rousseau:

> If it was a matter of hunting a deer, everyone well realized that he must remain faithful to his post; but if a hare happened to pass within reach of one of them, we cannot doubt that he would have gone off in pursuit of it without scruple…. (Rousseau 1761: 111)

- The idea is that in this game there is a big payoff to collaborative behaviour, but there is also some temptation or propensity to defect, which has its rewards for the defector, but not for the abandoned cooperator.

Risk and Payoff Dominance

- In the Stag Hunt, there are two Nash equilibria: {Stag, Stag} and {Hare, Hare}. Unlike Rich Friend, merely discovering a Nash equilibrium doesn't solve this game. How do we choose?
- {Stag, Stag} is **payoff dominant**: it is Pareto-superior to all other Nash equilibria.
- {Hare, Hare} is **risk dominant**: 'A player who chooses to hunt hare runs no … risk, since his payoff does not depend on the choice of action of the other player' (Skyrms 2001: 32).
- If one is confident of the other player's rationality, one ought to hunt stag: that is the Nash equilibrium with the highest payoff for both agents.
- But if one is unsure of the other – if one thinks them flightly, or subject to short-term tempation – then one ought to bear risk in mind, and choose the act that leads to the risk dominant equilbrium.
- If **1** is certain that **2** will hunt stag, so should **1**; likewise, if either suspects the other will hunt hare (either by mistake or by defecting from the cooperative arrangement), they should hunt hare also.

William James on Trust in Society

- If these players can **trust** one another, they can get their most desired outcome. If they are suspicious of the other, they should act so as to secure the hare – the problem there of course is that act **guarantees** there will be no trust.

> A social organism of any sort whatever, large or small, is what it is because each member proceeds to his own duty with a trust that the other members will simultaneously do theirs. Wherever a desired result is achieved by the co-operation of many independent persons, its existence as a fact is a pure consequence of the precursive faith in one another of those immediately concerned. A government, an army, a commercial system, a ship, a college, an athletic team, all exist on this condition, without which not only is nothing achieved, but nothing is even attempted. (James 1896: §9)

- The Stag Hunt illustrates James' concerns well – the central issue is how to ensure that the lack of faith of the **risk-averse** doesn't prevent the socially optimal (and Pareto optimal) outcome from being rationally chosen.

The Prisoner's Dilemma

> Tanya and Cinque have been arrested for robbing the Hibernia Savings Bank and placed in separate isolation cells. Both care much more about their personal freedom than about the welfare of their accomplice. A clever prosecutor makes the following offer to each. 'You may choose to confess or remain silent. If you confess and your accomplice remains silent I will drop all charges against you and use your testimony to ensure that your accomplice does serious time. Likewise, if your accomplice confesses while you remain silent, they will go free while you do the time. If you both confess I get two convictions, but I'll see to it that you both get early parole. If you both remain silent, I'll have to settle for token sentences on firearms possession charges. If you wish to confess, you must leave a note with the jailer before my return tomorrow morning.' (Kuhn 2019)

- A simpler version: *Each player decides, in isolation and simultaneously, whether they will receive 10 dollars or the other player will receive 20 dollars.*

The Prisoner's Dilemma Game

- This puzzle has something like the following pay-off matrix (Reiss 2013: 56):

Table 6: Prisoner's Dilemma

Tanya↓ / Cinque→	Stonewall	Confess
Stonewall	(2, 2)	(0, 3)
Confess	(3, 0)	(1, 1)

- The **unique Nash equilibrium is that each confesses**.
  - If someone stonewalls, the other will regret not confessing.
- The problem: the equilibrium strategies {Confess, Confess} are jointly not very attractive.
- They could do better by acting **cooperatively** with each other by staying silent – indeed, {Stonewall, Stonewall} is a Pareto-improvement on the Nash equilibrium, and is Pareto optimal.
  - Of course, the {Stonewall, Confess} and {Confess, Stonewall} strategies are also Pareto optimal, but are not a Pareto improvement on the Nash equilibrium, so we don't discuss them.

Instability of Cooperation

- In Stag Hunt, cooperation is **fragile**. In the Prisoner's dilemma, it seems rationally unobtainable:

> one might say that a PD is a game in which a 'cooperative' outcome obtainable only when every player violates rational self-interest is unanimously preferred to the 'selfish' outcome obtained when every player adheres to rational self-interest. (Kuhn 2019: §3)

- So why can't players just **agree to cooperate**?
- Here the Nash equilibrium gets some bite: for if the players cooperate, both will regret not defecting, given what the other player did: 'unless there is external enforcement … the incentives are strong not to honour the agreement' (Reiss 2013: 64).

> This influential argument runs as follows: Nash Equilibria are recommended by being the only strategy combinations on which the players could make self-enforcing agreements, i.e., agreements that each has reason to respect, even without external enforcement mechanisms. (Risse 2000: 366)

Self-Enforcement and Rational Decision

Table 7: A non-self-enforcing Nash equilibrium (Risse 2000: 366)

1↓ / **2**→	Left	Centre	Right
Top	(4,6)	(5,4)	(0,0)
Middle	(5,7)	(4,8)	(0,0)
Bottom	(0,0)	(0,0)	(1,1)

- In this game, {Bottom, Right} is the Nash equilibrium (any strategy other than Bottom or Right, and **1** or **2** will have regrets).
- But they should have regrets anyway, since if **1** had played any stratey other than Bottom, they would have done better if **2** has played any strategy other than Right.
  - Think about things this way: playing Not-Right pays off at least 4 for **2**, and playing Not-Bottom pays off at last 4 for **1**. So why can't mutually rational agents see that the 'collapsed' game is like modified Mozart or Mahler, and thus opt for a Pareto improvement, which they get regardless of how {Not-Bottom, Not-Right} is implemented (Reiss 2013: 64)?

Collective Action

- Many people have found Stag Hunt and Prisoner's dilemmas present in problems around **public goods** (lecture 10): those that are non-excludable (access uncontrolled once the good exists) and non-rivalrous (my use doesn't diminish availability).
  - E.g., national defence, road safety, or clean air.
- Consider the **tragedy of the commons**:

> Picture a pasture open to all. It is to be expected that each herdsman will try to keep as many cattle as possible on the commons. … the rational herdsman concludes that the only sensible course for him to pursue is to add another annimal to his herd. And another; and another.… But this is the conclusion reached by each and every rational herdsman sharing a commons. Therein the tragedy. Each man is locked into a system that compels him to increase his herd without limit—in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest in a society that believes in the freedom of the commons. (Hardin 1968: 1244)

The Tragedy of the Commons

Table 8: A tragedy of the commons (Kuhn 2019: §5)

1↓ / 2→	Moderate	Overstocking
Moderate	(2, 2)	(0, 3)
Overstocking	(3, 0)	(1, 1)

- In this game, each player can choose to pursue moderate usage of the commons, or overstocking it. Of the two Nash equilibria, the moderate usage one is better for both. But if they end up there, a lot of common grass goes unused, and overstocking is **individually better** – a short term windfall to the first mover who can get their herd onto the commons quickly. So it looks rational to end up with the despoiled commons.
  - This is generally an n-player game, for large n.
- So phrased, this is a prisoner's dilemma. But it may turn into a Stag Hunt, as the commons is despoiled and the advantages of overstocking diminish.
  - That doesn't mean cooperation is easier to establish; a population of longstanding non-cooperators might well suspect one another of being untrustworthy.

Coordination and Conflicting Interests

- So far we've seen cases where
  1. There is a best outcome for all, and it is a Nash equilibrium (original Mozart-or-Mahler).
  2. There is a best outcome for all, but it is not a Nash equilibrium (Prisoner's dilemma).
  3. There is a best outcome for all, but there is more than one Nash equilibrium (Stag Hunt).
  4. There is no best outcome for all, and no Nash equilibrium (Love and Loathe).
- We can also imagine a case where there is no best outcome for all but there is a Nash equilibrium:

Table 9: Never drink alone (aka 'Battle of the sexes')

1↓ / 2→	Beer	Wine
Beer	(2, 1)	(0, 0)
Wine	(0, 0)	(1, 2)

How can we coordinate?

- In Never Drink Alone, the players want to drink the same thing as each other, but each wants a different drink. How will they ever reach one of the Nash equilibria?
- The assumption that both players are rational and have common knowledge of rationality gives trouble, because the symmetry of the pay-offs makes **coordination** difficult (Peterson 2017: 245–46).
  - If opting for beer is **1**'s best strategy, then **2** will opt for wine, for the same reason. Likewise if **1** should opt for wine, **2** will opt for beer.
- To break the symmetry we could use **probabilities**: maybe beer is just more available, so it will be more likely that each player stumbles across some beer to drink. Or if we knew that **2** is such a fanatical wine drinker that they will play the Wine strategy no matter what.
  - Both of these involve again essentially giving up on the game theoretical approach – the second doesn't even have rational players!

Coordination from Self-Enforcement

> A self-enforcing agreement is one that provides incentives for the agents to stick to it even in the absence of external enforcement mechanisms. (Risse 2000: 368)

- Considerations of risk, just like those in Stag Hunt, might provide incentives that secure coordination even without agreement or a Nash Equilibrium.
- If we adopt the Risse definition, then this game is one where players have incentives to cooperate despite having better options if they do. But if they've agreed on the compromise, they have some reason to think they will stick to it, and that might be enough to support the otherwise fragile cooperative outcome.

Table 10: Don't risk defecting (Risse 2000: 368)

1↓ / 2→	Defect	Cooperate
Defect	(0, 0)	(4, 2)
Cooperate	(2, 4)	(3, 3)

A Clash With Decision Theory?

Israel/Egypt War Revisited

- Recall the Israeli/Egypt war example from lecture 3.
- If we now treat this as a strategic interaction, where Egypt is deciding on a military strategy and Israeli is deciding on an occupation strategy, we get the following pay-offs:

Table 11: Israel/Egypt revisited

Israel↓ / Egypt→	War	Peace
Occupy	(1, 1)	(3, 0)
Withdraw	(0, 3)	(2, 2)

- This is a Prisoner's dilemma!
  - The unique Nash equilibrium is {Occupy, War}, but again, there is a Pareto improvement available, namely, {Withdraw, Peace}.
- Applying game theoretic reasoning suggests the non-cooperative outcome is uniquely rational.

Two Rationalizations of Decision

- There is an immediate concern: as theories of rational choice, do game theory and decision theory apply to the same choice situations – and if they do, what is the guarantee that they provide compatible explanations?
- In our earlier treatment of the Israel/Egypt, we suggested that if we think that Israel's choice is **correlated with** Egypt's choice (Bar-Hillel and Margalit 1972: 296–97) – i.e., that occupation correlates with war, and withdrawal correlates with peace – then it can be rational to adopt the preferred bundle of strategies.
  - In effect, the correlations make the {Occupy,Peace} and {Withdraw,War} outcomes **negligible**, and the choice then is easy.
- But then it looks like we can rationalize both cooperation and defection; how then does rational choice theory **explain** actual behaviour, if it could explain either?

Parallel or Overlapping Rationality?

- Perhaps game theory may not **compete** with rational choice theory; it may be thought to apply in cases where we lack probability information.
- Perhaps this is forced on us in strategic interactions; some have argued that assigning probabilities to outcomes of choices – as we did in giving a decision-theoretic model of Israel-Egypt – is **incompatible** with seeing yourself and other players as deliberative agents, and so would argue that the tools of decision theory simply do not apply to strategic games (Levi 2007; but see Hájek 2016).
- Alternatively, standard approaches to decision theory might be thought to apply only where we have act-outcome independence, and game theory applies where we do not.
- Perhaps game theory and decision theory overlap, but actually give the **same verdicts**, properly understood.
- Perhaps we ought to adapt our approach to Israel-Egypt to the standard PD: : if the other prisoner is very much like me, a **replica** (Kuhn 2019: §7), I can rationally stonewall, because my choice then is between us both stonewalling and us both confessing.

Leaving Game Theory Behind

- This idea – play the probabilities of what the other player will do – **leaves game theory behind**: this is really just maximising expected utility with act-dependent probabilities.
- If players also have probability information about **likely strategies**, they don't need to reason strategically.

> So what is the point of identifying the Nash equilibria? Why not treat (non-cooperative) nonzero-sum games as individual decisions under risk? (Peterson 2017: 243)

- Perhaps the idea is that reasoning strategically **tells you** the likely probabilities. But does it? If I've got good evidence that my twin and I will act alike, then I know that we are rationally choosing between only {Confess, Confess} and {Stonewall, Stonewall}.
- Applying individual rational choice theory however might also involve the complexities of causal versus evidential decision theory: confessing dominates, and since my act doesn't **cause** my replica to act, I ought to confess.
  - Prisoner's dilemma might be a Newcomb Problem, 'or rather, two Newcomb Problems side by side, one per prisoner' (Lewis 1979: 235; Bermudez 2013).

Non-simultaneous Games

Dynamic games, Repeated Games

- Maybe some of the artefacts of our analysis arise because we are considering too limited a class of games.
  - Very few games involve a simultaneous and irrevocable choice of strategy!
- Many games might be played **repeatedly**, and that allows strategies for each repetition to possibly diverge from the rational strategy in one-shot games.
  - For example, repeated Stag Hunters might be able to **build trust** in cooperation through repeated encounters with one another, thus making the optimal equilibrium possible.
- Strictly speaking, such evidence about the players of repeated games gives information that goes beyond the available strategies and outcomes. But we can treat repeated games as single temporally-extended games with a larger space of strategies, which opens up new possibilities for analysis – and new difficulties, unfortunately.

Repetition and Cooperation

- In the prisoner's dilemma, cooperation is difficult to secure because the 'sucker's pay-off' for stonewalling is both a severe deterrent and a penalty for not foreseeing the rational choice of your partner in confessing.
- Note however that if you both stonewall, you get a desirable pay-off; and if you play **repeatedly**, you can establish a relationship that rewards achieving this pay-off, if each of you can foresee that the net gain from cooperating over time outweighs the foregone gain of taking advantage of another's cooperation:

> The main idea behind the theory of repeated games is that if the game is played repeatedly then the mutually desirable outcome in which [cooperation] occurs in every period is stable if each player believes that a defection will terminate the cooperation, resulting in a subsequent loss for him that outweighs the short term gain. (Osborne and Rubinstein 1994: 133)

Tit-for-Tat

- Consider these strategies in an **indefinitely repeated** prisoner's dilemma [Reiss (2013), p. 60:

Defect	Always defect (confess)
Cooperate	Always cooperate (stonewall)
Tit-for-tat	Cooperate on the first round of the repeated game; on round $n + 1$ do whatever the other player did on round n.

- Defect is the Nash equilibrium strategy in the one-shot game; each round, Defectors end up with no less than 2 utiles, and may end up with close to 4 on average if they play lots of Cooperators.
- Cooperators end up with no more than 3 utiles, and may end up with close to 1 on average if they play lots of Defectors.
- Tit-for-Tat plays like a Cooperator against Cooperators, and a Defector against Defectors. It typically does better than either, taking advantage of the **excess utility generated by mutual cooperation over mutual defection**.

Evolutionary Game Theory (Peterson 2017: §12.5)

- If we suppose that the strategies played are not rationally chosen but rather **genetically heritable**, and we suppose that the pay-offs correlate with **reproductive success**, we have **evolutionary game theory**.
- Here, a population consists of players with fixed strategies who interact with each other with certain probabilities.
- The key idea is of an **evolutionarily stable strategy** (Maynard-Smith 1982), which is a strategy that provides better outcomes for those who play it when played against itself than when played against any other strategy, and which is the best strategy to play against itself.
  - Such a strategy is stable, since it can never be **driven to extinction** in the population – if the numbers who play it get low, it encounters mostly strategies when it does better; if the numbers who play it get high, it encounters mostly itself, and it still does better.
- The **key result** is that cooperation can be an ESS. In a version of the **ultimatum game**, Skyrms shows that the one-shot irrational strategy of 'offer half' can emerge to take over the whole population (Skyrms 1996).

Backward Induction

- What if the game is repeated a **known finite** number of times?
- The situation changes drastically: cooperation does not emerge rationally, as can be seen by the **backward induction** argument (Reiss 2013: 60; Hausman, McPherson, and Satz 2017: 280–81).
- Suppose a PD is to be played 100 times.
  - On the last round, rational players know this is the last round. There is **no incentive** to cooperate for future reward; this is a one-shot game, and rational players defect.
  - On the second last round, rational players **know that on the last round everyone will defect**. So they know cooperation on the second last round will **not be rewarded**; so they should defect on that round too.
  - Likewise for every other round: everyone will rationally **always defect**, from the very first round.
- Players who reason this way end up with an average pay-off of ~2 on each round; players who treat 100 as if it is infinity average a pay-off of ~3 over each round.

Playing the Irrational

- **Given that the rational strategy in a finitely repeated prisoner's dilemma is to defect, what should Jill do if the person she is playing against, Jack, cooperates?**

> Jack's first move decisively refutes Jill's view of the game. It demonstrates that [there cannot be common knowledge of rationality]. But what if Jack's move is motivated by his knowing that making a cooperative move will post this perplexing problem for Jill and that such a move may thus induce Jill to cooperate in order to take advantage of Jack's apparent 'irrationality'! … It is hard to tell a convincing story of how a player should work out a suitable strategic response to contingencies that ought not to arise if all the players are rational and well schooled in game theory. (Hausman, McPherson, and Satz 2017: 282)

Further Topics

- This has barely scratched the surface of game theory and its significance, particularly for the rationality of cooperative social behaviour (Skyrms 1996).
- We haven't considered games where players choose their strategy over time rather than all at once; here the idea of **binding** oneself to a strategy is important (Hausman, McPherson, and Satz 2017: 283–85).
  - Consider here **nuclear deterrence**: the rational response to a first strike is not to retaliate (what's the point?), but deterrence requires that one make a credible threat of making the irrational response (Lewis 1986)!
- Nor have we considered extensively games where the players can bargain, and secure prior **agreements** to trust one another (Hausman, McPherson, and Satz 2017: §14.5).
  - Obviously such agreements can make securing the desired outcome in Stag Hunt and the PD much simpler, and they've also been used in solving problems about how to allocate scarce resources between groups with competing interests: clearly important for issues of distributive justice.

References

References

Bar-Hillel, Maya and Avishai Margalit (1972) 'Newcomb's Paradox Revisited', *The British Journal for the Philosophy of Science* **23**: 295–304. doi:10.1093/bjps/23.4.295.

Bermudez, J L (2013) 'Prisoner's Dilemma and Newcomb's Problem: Why Lewis's Argument Fails', *Analysis* **73**. doi:10.1093/analys/ant034.

Hájek, Alan (2016) 'Deliberation Welcomes Prediction', *Episteme* **13**: 507–28. doi:10.1017/epi.2016.27.

Hardin, Garrett (1968) 'The Tragedy of the Commons', *Science* **162**: 1243–48.

Hausman, Daniel, Michael McPherson, and Debra Satz (2017) *Economic Analysis, Moral Philosophy, and Public Policy*, 3rd edition. Cambridge University Press.

James, William (1896/1956) 'The Will to Believe', in *The Will to Believe and Other Essays in Popular Philosophy*: 1–31. Dover.

Kuhn, Steven (2019) 'Prisoner's Dilemma', in Edward N Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/prisoner-dilemma/.

References (cont.)

Levi, Isaac (2007) 'Deliberation Does Crowd Out Prediction', in T Rønnow-Rasmussen, B Petersson, J Josefsson and D Egonssson, eds., *Hommage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz*.

Lewis, David (1979) 'Prisoners' Dilemma Is a Newcomb Problem', *Philosophy and Public Affairs* **8**: 235–40.

Lewis, David (1986) 'Buy Like a MADman, Use Like a NUT', *QQ* 5–8.

Maynard-Smith, John (1982) *Evolution and the Theory of Games*. Cambridge University Press.

Osborne, Martin J and Ariel Rubinstein (1994) *A Course in Game Theory*. MIT Press.

Peterson, Martin (2017) *An Introduction to Decision Theory*, 2nd edition. Cambridge University Press.

Reiss, Julian (2013) *Philosophy of Economics*. Routledge.

Risse, Mathias (2000) 'What Is Rational about Nash Equilibria?', *Synthese* **124**: 361–84. doi:10.1023/a:1005259701040.

Rousseau, Jean-Jacques (1761/1984) *A Discourse on Inequality*, Maurice Cranston, trans. Penguin.

Skyrms, Brian (1996) *Evolution of the Social Contract*. Cambridge University Press.

References (cont.)

Skyrms, Brian (2001) 'The Stag Hunt', *Proceedings and Addresses of The American Philosophical Association* **75**: 31–41.