THE PRISONER'S DILEMMA PARADOX: RATIONALITY, MORALITY, AND RECIPROCITY

Rory W. Collins

This article examines the prisoner's dilemma paradox and argues that confessing is the rational choice, despite this probably entailing a less-than-ideal outcome.

The Prisoner's Dilemma

Imagine two suspected criminals, Andrea and Bryan, are arrested for importing illegal weapons and placed in separate cells. A police officer enters Andrea's cell and informs her there is not enough evidence to convict the pair as it stands. If both she and Bryan remain silent, they will be charged only with a minor offence and sentenced to one year each in prison. If one of them confesses while the other stays silent, the confessor will be released for turning state's witness and their accomplice will serve a full decade in jail. If they both confess, each of them will receive five years behind bars. Bryan is being told the same information in his cell. The officer tells Andrea to decide what to do before she and Bryan are interviewed separately that evening.

The above situation describes a classic prisoner's dilemma. Mark Sainsbury, in his portrayal of the paradox, clarifies two assumptions inherent to the predicament (Sainsbury 2009: 83):

(1) Each prisoner is only concerned with getting the minimum sentence for themselves.

doi:10.1017/S1477175621000464
Think 61, Vol. 21 (Summer 2022)

Table 1. Possible outcomes of the two prisoners' decisions

		Bryan	
		Confess	Stay Silent
Andrea	Confess	(5, 5)	(0, 10)
	Stay Silent	(10, 0)	(1, 1)

(2) Neither has any information about the likely behaviour of the other except that (1) holds for them and that they are a rational agent.

The potential outcomes can be expressed in the above table, where the numbers in each cell represent the years Andrea and Bryan will respectively serve:

What would be rational for Andrea to do? Looking at Table 1, it seems clear on the face of it that she should confess since whatever Bryan does, she will minimize her sentence by confessing. If Bryan stays silent, she will get off scot-free, saving herself a year in jail. If Bryan confesses, Andrea will serve five years: a somewhat disappointing outcome but preferable to the ten years she would receive if she stayed silent. This persuasive line of reasoning is based on the dominance principle, which holds that if one choice gives an equivalent or better outcome to all other available choices – that is, it *dominates* the range of alternatives – then this choice is the rational decision. However, since Bryan is rational too, he will presumably reason the same; hence both prisoners will confess and secure themselves five years in jail. By acting in this apparently rational way, each ends up worse off than if they had both kept silent.

Several philosophers have advanced a case for staying silent on rational grounds called the symmetry argument (Bicchieri and Green 1997). The general idea is that Andrea knows she and Bryan are both rational agents, so whatever choice she arrives at, Bryan will also reach, since two rational agents will reason symmetrically. This reduces

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

the four possibilities to just two: both confess or both stay silent. Of these, the latter is clearly preferable, so Andrea should remain silent, confident that Bryan will too.

Sainsbury raises (and later dismisses) a slightly different form of the symmetry argument based on the principle of maximizing expected utility. When applied to the present case, this argument holds that Andrea need not be certain that Bryan's decision will match hers, only that there is a sufficiently high probability that their choices will match to justify silence. If Andrea confesses, then, since both she and Bryan are rational agents, there is a good chance Bryan will also confess. Let m denote the probability that their choices match. This means the probability he will remain silent, given Andrea confesses, is $1 - m$ since there are only two alternatives. So, if Andrea confesses, her expected years in prison can be calculated by taking the sum of the two outcomes multiplied by the probability of their occurrence, as follows:

$$E(X_C) = (5 \times m) + (0 \times [1 - m]) = 5m$$

If Andrea stays silent, then again, there is a high probability, m, that Bryan will also stay silent, giving a $1 - m$ probability of him confessing. Her expected jail time if she remains silent is:

$$E(X_S) = (1 \times m) + (10 \times [1 - m]) = 10 - 9m$$

If opting for silence is to minimize Andrea's time in prison, the expected outcome of staying silent must be less than the expected outcome of confessing. That is:

$$E(X_S) < E(X_C)$$

$$10 - 9m < 5m$$

$$\sim 0.7143 < m$$

Comparing the expected outcomes of Andrea's two possible choices appears to show that staying silent is in her best

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

interests so long as she has reason to think that Bryan's decision will match hers with a 71 per cent chance or greater.[1]

The paradox of the prisoner's dilemma is now clear. There are plausible arguments for confessing and staying silent, yet both cannot be true since together they recommend that Andrea confess *and* stay silent: a blatant contradiction.

This paradox can be resolved by showing that the case for silence is flawed. Suppose both prisoners reason in the way prescribed by the symmetry argument. Andrea, believing that Bryan's decision will match hers, can see that if they both remain silent, they will both serve a year in jail. At this point, she should realize that if she now switches to confessing, she will save herself this year behind bars. Her decision to stay silent would be rational only if it were stable when considered against alternative choices; here, it is not. Andrea and Bryan's choices are causally independent of one another, so there is nothing about her choosing to confess or stay silent that necessitates Bryan will do so too. The expected utility calculations which generate a 71 per cent threshold for matching are tainted by the false assumption that one prisoner's decision is conditional on the other's, when in fact, as Sainsbury notes, their respective choices are separate events. Thus, the dominance argument remains persuasive, and confession appears the rational choice for both Andrea and Bryan, even though this is likely to lead to each of them spending longer in jail than if they had both remained silent.

There is something unsettling about the rational pursuit of self-interest leading each prisoner to a less-than-optimal result. Of course, there are many familiar circumstances where acting rationally can lead to a poorer consequence than acting irrationally. Folding on a risky poker hand may be rational, but a lucky turn of cards can sometimes lead to a significant payoff. In the case of the prisoner's dilemma, though, the outcome is not simply a matter of chance. The five years in jail that Andrea and Bryan receive is a foreseen and predictable consequence of their apparently

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

rational decisions, which is what makes the outcome seem inadequate.

This result has implications for the common notion of rationality, often conceived as requiring that rational agents act in ways expected to bring about the greatest satisfaction of their interests. The prisoner's dilemma, it seems, provides a counterexample to this definition since both prisoners knowingly produce a worse outcome than that which two irrational prisoners would reach under the same circumstances.

Morality

Perhaps something is wrong with the way the dilemma itself is constructed, and this is what generates the unsatisfactory result. There are two crucial conditions given in the thought experiment: (1) both prisoners are entirely selfish and (2) both prisoners are rational agents. But maybe the coexistence of these is impossible. If a person is rational, this may entail that they are not purely self-interested. Many philosophers, including Aristotle, Immanuel Kant, and John Stuart Mill, have claimed that their respective moral theories are rationally defensible. If a connection between rationality and altruism can be established, (1) and (2) must contradict, and this hidden contradiction may be what leads to the unacceptable conclusion of the dilemma.

Objecting to the prisoner's dilemma on the grounds of moral rationalism has received minimal attention in the literature. László Mérő claims that since the arguments for confessing and silence are equally sound, yet together lead to a contradictory conclusion, the dilemma must be logically impossible (Mérő 1998: 31). Mérő does not, however, suggest that this arises from the prisoners being required to act in both a self-interested and a rational manner. David Gauthier asserts that the prisoner's dilemma illustrates why we ought to accept a form of moral contractarianism since everyone is better off agreeing to cooperate over the long term (Gauthier 1986: 82, 169). Rational individuals, he

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

argues, should 'constrain' their attempts at utility-maximization and cooperate when pitted against other rational agents in prisoner's dilemma-type situations since doing so will provide them with a preferable outcome after repeated interactions. This fails to justify staying silent in a one-shot dilemma, though, since it can be assumed in such a thought experiment that the prisoners will not find themselves in identical circumstances again. Hence, cooperating on the egoistic grounds Gauthier suggests does not override the dominance argument.

Kant is among the most ardent defenders of the claim that rationality entails morality and proposed the categorical imperative as a moral panacea discoverable to rational agents *a priori*. In Kant's view, what makes humans unique is that they possess the ability to reason autonomously. This gives rise to a fundamental human dignity – the 'good will' – which alone is intrinsically valuable. The value of anything else is conditional on its use. Happiness, for example, is only good to the extent that a person gains it deservingly. If a person derives happiness from torturing others, their happiness is not warranted and therefore not good. All actions must, according to Kant, be performed out of duty towards a moral law which applies under all circumstances. Rational agents ought to respect the autonomy of other rational agents by not using them solely as means and must also be willing to act in ways that can be generalized across all circumstances. Fittingly, Kant's first formulation of the categorical imperative is to 'act only in accordance with that maxim through which you can at the same time will that it become a universal law' (Kant 2002: 37).

Kantian ethics requires that Andrea's decision to confess or stay silent must be motivated by a duty to abide by this imperative. Lying is not viable since doing so could not be universalized as a moral law. Both her and Bryan's decisions must, therefore, be based on whether they did in fact commit the crime; self-interested motives need not feature in the decision-making process. The prisoner's dilemma would cease even to be a dilemma, much less a paradox.

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

However, Kant's argument for moral rationalism is unconvincing. Many have argued that rationality alone is insufficient to motivate moral behaviour and that emotions are a necessary driving force behind moral acts. Arthur Schopenhauer claimed that Kant's theory provided no rational basis for what should count as a universal law and thus allowed for egoistic interpretations of the categorical imperative based on what individuals desire to be universalized. Compassion for others must be acknowledged as an essential component of morality, which a purely rational approach neglects. G. W. F. Hegel remarked that some moral maxims present contradictions when universalized and hence must be deemed immoral by Kantian ethics. For instance, if everybody donated to the poor, poverty would cease to exist, meaning there would be no poor people to donate to, and thus a contradiction. Kant's argument that rationality requires his form of deontological ethics is less persuasive than it first appears.

Most contemporary philosophers are equally unconvinced by Kant's moral theory. A recent study found that just 25.9 per cent of professional philosophers endorse some variety of deontology. Alternative systems fare no better, though, with only 23.6 per cent supporting consequentialism and 18.2 per cent supporting virtue ethics (Bourget and Chalmers 2014). No rational defence of morality has yet been proposed which has convinced the majority of experts in the field. Though it is possible such an argument might one day be made, a firm connection between rationality and morality has not yet been established. The prisoner's dilemma, then, appears not to rely on contradictory premises. It is quite possible to be rational and purely self-interested.

This leaves us with the somewhat depressing result that Andrea and Bryan will both confess and thereby land themselves with an additional four years in jail each than if they had remained silent. Such an unfortunate conclusion occurs through no fault in their reasoning, however. It is merely due to each prisoner being unable to determine

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

what the other will do. Opting for silence cannot influence the other prisoner's decision, so the dominance argument still applies from Andrea's and Bryan's individual standpoints. Both prisoners being rational and self-interested leads to the surprising, but not paradoxical, conclusion that Andrea and Bryan should both confess.

Reciprocity

Analogies can be drawn between the abstract thought experiment and real-life interactions. In many circumstances, the decisions of two or more self-interested agents may elicit a consequence that is worse for each of them than what could be achieved through cooperation. When several employees are required to work together on a task, the best outcome for any individual is to contribute little and leech off the efforts of his co-workers. Likewise, if one petrol station, located near a competitor, lowers its price to gain more customers, it will maximize its profit only if the rival business does not follow suit. On an international scale, consider the CO_2 emission restrictions that are required over the coming years to prevent climate change reaching catastrophic levels. For each individual nation, the dominant strategy is to continue polluting and allow other countries to bear the burden of reducing emissions, yet if too many countries do this, everyone will be worse off. Competing pursuits of self-interest, it seems, may lead to poorer outcomes for all.

There is a crucial difference between these examples and the one-shot prisoner's dilemma. In all three cases above, the decision an individual, business or nation makes at one time may have an impact on the decisions made by others in future interactions. This is not the case in a one-shot dilemma. The slack employee may be overlooked for a promotion. The petrol station may have its low prices equalled by its competitor. Nations who pollute the planet may find themselves faced with economic sanctions or boycotts. All three situations can be modelled more

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

accurately as iterated prisoner's dilemmas, where two parties decide whether to cooperate or defect over successive trials. Computer simulations show that in such dilemmas, the optimal strategy is 'tit-for-tat', which involves cooperating for the first trial, then mirroring the opponent's last decision for each trial thereafter (Axelrod 1984: 31). Similar to Gauthier's notion of constrained maximization, this strategy allows one to maximize the benefits of playing against cooperative opponents without risking exploitation when pitted against those less inclined to cooperate. Despite a lack of rational arguments for altruism, the intuitive 'golden rule' to act towards others as you would have them do to you appears to benefit each self-interested party in an iterated prisoner's dilemma. In many of the real-life cases that approximate to these dilemmas, rational self-interest may not necessarily lead to a poorer outcome since cooperation can be justified on egoistic grounds alone.

Conclusion

The one-shot prisoner's dilemma appears to provide a serious challenge to rationality since two rational prisoners would elicit a worse outcome than two irrational prisoners, despite their being able to foresee this as a consequence of making the dominant choice. However, this occurs due to each prisoner being unable to control the other's actions, not because of a flaw in their reasoning. The symmetry argument for silence is unconvincing since the prisoners' decisions are causally independent; Andrea staying silent cannot influence Bryan to do the same. Likewise, the rational morality argument against the logical possibility of the dilemma is unpersuasive since neither Kant nor other philosophers have managed to demonstrate conclusively that rationality entails morality. The real-life applications of the dilemma are vast, and, in many cases, cooperation can be justified without invoking altruism since acting selfishly risks this being reciprocated in the future, leading to a worse outcome for all involved. Ironically, self-centredness

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

may be among the factors which can motivate cooperative behaviour.[2]

Rory W. Collins recently completed a Master of Teaching and Learning in Secondary Education at the University of Canterbury, New Zealand. He now works as a high school mathematics teacher. rorycollins97@gmail.com

Notes

[1] Expected utility calculations for different sentence lengths are included in an appendix.

[2] Many thanks to Joshua Leota for several engaging conversations on this topic, and to Stephen Rowe and Zhuo-Ran Deng for their helpful comments on earlier drafts.

References

Axelrod, R. (1984) *The Evolution of Cooperation* (New York: Basic Books).

Bicchieri, C. and Green, M. S. (1997) 'Symmetry Arguments for Cooperation in the Prisoner's Dilemma', in Ghita Holmström-Hintikka and Raimo Tuomela (eds.) *Contemporary Action Theory*, vol. 2 (Dordrecht: Kluwer Academic Publishers), 229–49.

Bourget, D. and Chalmers, D. J. (2014) 'What Do Philosophers Believe?', *Philosophical Studies* 170.3: 465–500.

Gauthier, D. (1986) *Morals by Agreement* (New York: Oxford University Press).

Kant, I. (2002) *Groundwork for the Metaphysics of Morals*, trans. Allen W. Wood (New York: Vali-Ballou Press).

Mérő, L. (1998) *Moral Calculations: Game Theory, Logic, and Human Frailty*, trans. Anna C. Gösi-Greguss (New York: Springer-Verlag).

Sainsbury, R. M. (2009) *Paradoxes*, 3rd edn (Cambridge: Cambridge University Press).

Appendix: Expected Utility Calculations

Table 2 shows the outcomes each prisoner may receive with the specific prison sentences replaced by the general terms *p*, *r*, and *s*.

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press

Table 2. Generalized possible outcomes of the prisoner's dilemma

		Bryan	
		Confess	Stay Silent
Andrea	Confess	(p, p)	(0, s)
	Stay Silent	(s, 0)	(r, r)

Consider the expected outcome for Andrea. Let the probability that her and Bryan's decisions match be denoted as m. Her expected years in prison for either confessing or staying silent are:

$$E(X_C) = (p \times m) + (0 \times [1 - m]) = pm$$

$$E(X_S) = (r \times m) + (s \times [1 - m]) = rm + s - sm$$

For staying silent to minimize Andrea's time in prison, the expected outcome of staying silent must be less than the expected outcome of confessing. That is:

$$E(X_S) < E(X_C)$$

$$rm + s - sm < pm$$

$$s < m(s + p - r)$$

$$\frac{s}{s + p - r} < m$$

So, for variants of the prisoner's dilemma with different prison sentence lengths, the above inequality must be satisfied to justify silence on the grounds of maximizing expected utility.

https://doi.org/10.1017/S1477175621000464 Published online by Cambridge University Press