GAME THEORY AND THE SELF-FULFILLING CLIMATE TRAGEDY

(forthcoming in *Environmental Values*)

Matthew Kopec
The Centre for Applied Philosophy and Public Ethics

Prepublication version as of 5 September 2016
Please do not quote without prior permission, since portions might change during copyediting.
Comments are very welcome: matthewckopec@gmail.com

Abstract: Game theorists tend to model climate negotiations as a so-called 'tragedy of the commons'. This is rather worrisome, since the conditions under which such commons problems have historically been solved are almost entirely absent in the case of international greenhouse gas emissions. In this paper, I will argue that the predictive accuracy of the tragedy model might not stem from the model's inherent match with reality but rather from the model's ability to make self-fulfilling predictions. I then sketch some possible ways to dispel the tragedy, including (1) recognising some ways the assumptions of the model fail, (2) taking seriously recent work suggesting that increasing greenhouse gas emissions is not in most nations' own self-interest, and (3) preferring alternative models like collective risk dilemmas, bargaining games, or cooperative models.

Key Words: Climate Negotiations, Climate Change, Game Theory, Tragedy of the Commons

1. Introduction

With the exception of a few very recent glimmers of hope, international climate change negotiations have been a disaster. Just to give one example, the Warsaw Climate Conference in 2013 ('COP19') utterly collapsed, with a coalition of developing nations simply walking away from the meeting. A group of key NGOs, including such important international players as Greenpeace, Oxfam International, and the World Wildlife Fund, also eventually refused to participate. Here is an excerpt from the letter of protest these NGOs penned:

> The Warsaw Climate Conference, which should have been an important step in the just transition to a sustainable future, is on track to deliver virtually nothing. In fact, the actions of many rich countries here in Warsaw are directly undermining the [United Nations Framework

> Convention on Climate Change] itself, which is an important multilateral process that must succeed if we are to fix the global climate crisis.[1]

The continued failure of these climate negotiations is, in many respects, exactly what game theorists have been predicting for decades. These scholars tend to model the negotiations as a competitive interaction between nation-actors, and the game theoretic model most have settled upon is the so-called 'tragedy of the commons.'[2] On the tragedy model, as I'll refer to it, it is in each nation's self-interest to abuse the commons, since they can each yield economic benefits by producing greater amounts of greenhouse gasses, while the costs due to the resulting changing climate are spread around to everyone. Given that the standard conditions needed to solve one of these commons problems are almost entirely absent in this particular case, many game theorists have predicted that the impending tragedy is inevitable. And with each climate meeting that collapses, and each climate treaty that is broken, the game theorists gain more confirmation for their model.

Or so it seems. I am actually not so pessimistic. It was noticed, long ago, that models in the social and behavioural sciences, and in economics in particular, often have the capacity to change the way the social world operates.[3] Some of these can cause the social world to better conform to the models than it would if the models never gained prominence. In other words, some such models give rise to self-fulfilling prophecies. It is possible that as economists and game theorists have settled upon using the tragedy of the commons model to analyse international climate negotiations, they have created one of these self-fulfilling prophecies. By spreading the prediction that each nation would attempt to defect from any collaborative agreement these theorists have, in turn, increased the likelihood of such defections. In this paper, I make this case, in the hopes of convincing the reader that there is some room for optimism. [4]

The plan of this paper is as follows. In Section 2, I examine how climate negotiations have been sold as a game-theoretic tragedy of the commons. In Section 3, I discuss one case in which economic models tend to be self-fulfilling,

---

[1] Available at http://wwf.panda.org/?212532/NGOs-Social-Movements-Walk-Out-Of-Warsaw-Talks (accessed 7 July 2016).

[2] For some examples see (Soroos, 1997; Ostrom et al., 1999; Barrett, 2003; Johnson, 2003; Sandler, 2004; Gardiner, 2006; Binmore, 2007; Harris, 2010; Bernauer, 2013). Some scholars model the problem as a prisoners' dilemma, which is actually a special case of the tragedy of the commons. So, I'll lump these scholars together with those who favour the tragedy model more explicitly. Gardiner presents some reasons to think the broader tragedy model fits the problem better (Gardiner 2011: 108-11).

[3] The literature on this phenomenon is extensive. For some classics see (Merton, 1948; Simon, 1954; Rosenthal and Jacobson, 1968; Callon, 1998; Soros, 2003; MacKenzie, 2008).

[4] I admit, at the outset, that I will not be able to make an absolutely airtight case in the coming pages. Doing so would take many years of careful (and difficult) sociological work. I only intend to make the possibility seem likely, which I hope will help redirect scholarly efforts toward areas that promise a bit more optimism.

namely, the case of economics education, and I argue that the widespread modelling of climate negotiations as a tragedy of the commons has many relevant similarities with other self-fulfilling cases in economics. In Section 4, I sketch some possible strategies for dispelling the self-fulfilling tragedy, if it proves to be one. I conclude in Section 5.

2. Modelling Climate Change as a Tragedy of the Commons

Informally speaking, a 'tragedy of the commons' is an economic phenomenon whereby it is in the individual best interest of every member of a community to increase their use of a common pool resource to the detriment of the common good. Here is a classic example (Lloyd, 1833). Say we have a community of farmers who collectively share a patch of land upon which they all graze their herds of goats. If a farmer adds a goat to her herd, then there is slightly less grass for each goat to graze on, meaning the goats each gain slightly less weight, produce slightly less milk, etc. The losses caused by this one farmer adding an additional goat are not paid solely by the farmer who added the goat—the losses are spread among all the herds collectively. So, while this farmer who adds a goat personally yields all of the benefits of doing so, the costs are spread throughout the population. The same rationale that holds for adding the first additional goat holds for the next, and the next, up to the point where adding another goat would harm the farmer's own herd so much that it would not be worth adding. And since each farmer will reason the same way, each will put more and more goats on the commons, perhaps well past the point of ruining it.

Game theorists can model this kind of situation mathematically. Take the following analysis, roughly following Binmore (2012: 29). Say we have ten farmers in our community, and each farmer gets a payoff per goat that she puts on the commons that is given by the function $m = e^{1-1/10n}$, where m is the monetary pay out (say in pounds sterling c.1833), n is the total number of goats from all farmers, and e is the natural constant (i.e., Euler's number).[5] So, if each of our ten farmers puts one goat on the commons, then each goat pays out $£e^0$, meaning that each farmer gets one pound for the goat. Since we have ten farmers, this means the commons is capable of producing ten pounds in wealth for our farming collective. To determine whether these farmers would be content with just a single goat in their herds, we would need to look more closely at the payoff functions, which I present below in Figure 1.[6]

[5] The specific mathematical details here are largely unimportant. They're just designed to give the payoff curves roughly the right characteristics.

[6] Since the figure on the horizontal axis is the average number of goats placed by the other farmers, one computes n in the payoff function simply by multiplying that figure by 9, i.e., the number of other farmers, and adding the last farmer's goats. I thank an anonymous reviewer for helping me improve this figure.

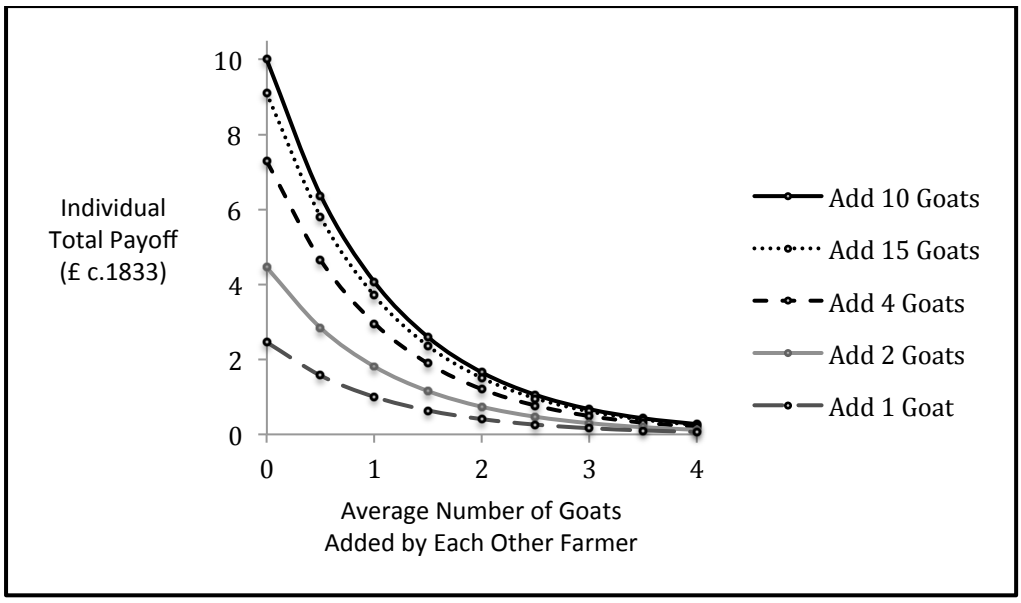

Figure 1

A few things are worth noticing. First, this farmer's total payoff is inversely related to the total number of goats on the common, i.e., it decreases as the average number of goats placed by the other farmers increases. But if we hold fixed the average number of goats the other farmers place on the commons, it is in this farmer's best interest to add two goats instead one, four instead of two, etc., until she has ten goats in her herd. It is not until she adds more than ten goats that her total payoff starts to actually diminish.[7] So how many goats should she, rationally speaking, place on the commons? One reasonable proposal is that she should put the number of goats on the commons that guarantees that she maximizes her payoff, *no matter what the other farmers decide to do*. (In game theoretic terms, she should play the 'Nash equilibrium'.) In this case, that means putting ten goats on the commons. Each of the other farmers should in turn reason the same way, also putting ten goats on the commons, and this leads to a very bad result. (So bad, it well off to the right of Figure 1.) Now the payoff per goat has been decreased to $£e^{-9}$, or £0.012. So, whereas the commons could have produced ten pounds in total wealth if they had all just added one goat to the pasture, their individual, rational self-interest would drive them to produce a mere 12 pence. This is the nature of the tragedy: when each individual acts in her own rational self-interest, it destroys the collective good.

So does international greenhouse gas emissions meet the conditions for a tragedy of the commons? When Garrett Hardin reintroduced the tragedy of the commons into the contemporary literature he isolated two characteristics, namely, that the benefits of abusing the commons go directly to the individual, while the costs of abusing the commons are spread throughout the population

[7] I used 15 goats in the figure just to make the decline more noticeable, but the decline in payoff starts with goat number 11.

(Hardin, 1968). This certainly seems to be met in the case of greenhouse gas emissions. Let us look at things from the perspective of one country, say, the United States. The US yields massive economic benefits from burning fossil fuels, by way of inexpensive energy production, heating, and transportation. If the US decides to burn higher carbon emitting (but cheaper) fossil fuels like coal over lower emitting (but more costly) fuels like natural gas, it yields additional benefits. And, although other greenhouse gases are often ignored in the discussion, the US economy further prospers when massive amounts of methane are released during its natural gas production[8] or through the burps of its roughly 40 million cattle.[9] All of these benefits go directly to the US economy, often measured as a boost in Gross Domestic Product (GDP). The costs of these emissions, on the other hand, are spread around the world. It is true that the US will pay serious costs if global temperatures rise to catastrophic levels, but the US will not be the only ones paying such costs. In fact, many have argued that the nations most responsible for our current levels of greenhouse gasses, i.e., the earliest industrialised nations in North America and Europe, are also those best suited to survive a rapidly changing climate (Gardiner, 2011).

The case of international greenhouse gas emissions also has a number of compelling parallels with our goat-tragedy model from above. Instead of farmers placing goats, we have nations emitting tonnes of carbon dioxide and methane. Instead of a grazing pasture, we have the carbon clearing capacity of the biosphere. And the payoff structure seems to act in a very similar way. Take the US again. Holding fixed the carbon emissions of all other nations, it is in the US's best interest to emit more carbon than less, up to the point where the next few tonnes could yield a net cost. But the benefits of emitting a certain amount of carbon decreases as the other nations also increase their emissions. In a pre-industrialised world, emitting a sizeable amount of carbon has little noticeable effect, since the biosphere can adjust. But emitting that same amount of carbon in a rapidly industrialising world means real costs, such as sea level rise, massive storms, crop failures, etc. Given these parallels, we can assume there is an emissions point for each nation that provides that nation optimal gain no matter what the other nations choose to do with their emissions. We can also assume that each nation emitting at this rationally self-interested level will lead to absolute disaster.[10]

Given the parallel with existing tragedy models, and given that Hardin's original conditions are met, it is not surprising that the tragedy model now dominates the literature on climate negotiations across various disciplinary perspectives. (See e.g., Soroos, 1997; Ostrom et al., 1999; Barrett, 2003; Sandler, 2004; Binmore, 2007; Harris, 2010; Gardiner, 2011; Johnson, 2011; Bernauer, 2013.) The International Panel on Climate Change (IPCC), in their 2014 report,

[8] https://www3.epa.gov/climatechange/ghgemissions/gases/ch4.html (accessed 7 July 2016).

[9] http://www.beefusa.org/beefindustrystatistics.aspx (accessed 7 July 2016).

[10] There are admittedly problems with the story here. I return to this point later.

put *high confidence* in the claim that 'climate change is a case of "the tragedy of the commons"'(211), and stated that 'Effective mitigation of climate change will not be achieved if each person or country acts independently on its own interests' (214). So it is safe to say that the model has really taken a hold in a range of fields working on climate negotiations.

Admittedly, not all problems that fit Hardin's two conditions are guaranteed to end in tragedy. Hardin himself presented the tragedy of the commons in order to argue for a certain solution to overpopulation, which he also saw as meeting the conditions. For Hardin, there were two main paths to solving a commons problem. First, a centralised authority, such as a state or local government, could maintain common ownership but somewhat restrict usage rights over those using the commons. This path coerces the individuals using the commons so that their own self-interest cannot drive them to overuse it. Second, the centralized authority could divide the commons up for private ownership, by giving the users individual rights to use their portion of the resource however they see fit. This path restructures the payoffs so that both the benefits *and the costs* of abusing the commons go directly to the individual in question. But neither of these solutions offers us much hope in the case of the climate commons. We lack an international, centralised authority that has the power to restrict a nation's usage of the commons or to grant a nation property rights over a portion of the climate commons (and then protect those rights).

The lack of a centralised authority obviously makes solving the climate commons problem more difficult, but Elinor Ostrom has shown that commons problems do not always end in tragedy without a centralized authority.[11] Ostrom's groundbreaking fieldwork uncovered numerous cases in which what she calls 'common pool resources' were effectively managed without the directly coercive means Hardin believed were necessary. Extrapolating a bit,[12] Ostrom and her colleagues isolated five conditions that had to be present in a community if the common pool resource was to be properly managed:

1) Resource use can be accurately monitored at low cost.
2) Resource use, technology, and economic and social conditions change slowly.
3) There is a dense social network among the actors using the commons.
4) Outsiders can be excluded from using the resource at low cost.
5) Users support monitoring and rule enforcement.

Unfortunately, every one of these conditions seems to fail in the case of greenhouse gas emissions. It is extremely difficult to accurately monitor each nation's net contribution of greenhouse gasses to the atmosphere, especially given the active debates over how to measure carbon sinks. And even if accurate monitoring were possible, it would not come at a low cost. The pace of technological change and industrialisation is rapidly changing social conditions

[11] See (Dodds, 2005) for additional examples.

[12] See (Ostrom et al., 1999; Dietz, Ostrom, and Stern, 2003; Ostrom, 2015) for the full details.

around the globe. The main actors in the climate commons are massively diffuse, to the point where it is even hard, at times, to tell who the main actors are (e.g., are multinational corporations also actors?). Outsiders, like the developing nations, usually cannot be 'excluded' from the commons without the use of military intervention, which comes at a high monetary cost (leaving the additional ethical costs aside). And as the weak monitoring and enforcement mechanisms of previous climate treaties suggest, some of the important players in climate negotiations do not genuinely support monitoring and enforcement.[13] As with Hardin's two conditions for solving a commons problem, all of Ostrom's conditions seem to fail for the climate commons as well.

So, if the tragedy of the commons is the right way for us to model the problem of international greenhouse gas emissions, then things look bleak. Moreover, as I mentioned earlier, the model seems to have much going for it. In particular, over two decades of negotiations have played out largely as the game theorists pressing the model would have expected, which suggests that the model has received a fair bit of empirical confirmation. We are left with only a few key options if we want to remain optimistic. First, we could contrive an alternative model that is just as well confirmed as the tragedy model yet lacks the pessimistic predictions. Second, we could try to argue that the case of greenhouse gas emissions actually doesn't meet the minimal conditions needed to make it a tragedy of the commons. Third, we could try to find some new set of conditions for solving a commons problem—conditions similar to Hardin's or Ostrom's but which can be met in the case of greenhouse gas emissions. Or, last, we could attempt to explain away the predictive accuracy of the tragedy of the commons. In this paper I will take this last path, although hints of the other paths will start appearing toward the end.

3. Self-Fulfilling Economic Models

Our typical understanding of how a scientific model gains confirmation relies upon the assumption that the predictions the model makes do not, themselves, have a causal effect on the phenomenon being modelled. Once we drop this assumption, it's possible for the confirmation relation to become erratic (Kopec, 2011). And ever since Robert Merton's classic essay (Merton, 1948), it has been well accepted that this assumption of predictive causal irrelevance, which is needed to keep the confirmation relation in order, fails in a massive way in the social and behavioural sciences.[14] Predictions in politics can elicit bandwagon and

[13] The Kyoto protocol is a key example. As Gardiner has aptly argued, even though the monitoring and compliance mechanisms built into the Kyoto treaty were much lauded at the time, a closer inspection suggests they were, in fact, deliberately made toothless (Gardiner 2011: 136-7). I feel it is safe to assume that this wouldn't have occurred if all the major players in the negotiations genuinely supported greenhouse gas monitoring and enforcement. I thank an anonymous referee for helping me clarify this point.

[14] These kinds of somewhat enigmatic causal effects have gone by a number of names in the literature, including self-fulfilling or self-frustrating prophecies or predictions (Merton 1948),

underdog effects that can change the outcomes of elections (Simon, 1954). Predictions in education can elicit the Pygmalion effect, where students perform in proportion to their teachers' expectations (Rosenthal and Jacobson, 1968). Predictions about how an institution would function best can affect how that institution functions (Ferraro, Pfeffer, and Sutton, 2005; Ferraro, Pfeffer, and Sutton, 2009). And economic predictions about the value of an asset or the solvency of a financial institution can directly affect the trading price of that asset or the solvency of that institution (Soros, 2003; Soros, 2013). So when the predictions of a social or behavioural scientific model come true, we need to be fairly certain that the predictions did not have a causal effect on the predicted outcome before we can safely infer that those predictions confirm the model that gave rise to them.

In fact, economics seems especially prone to this kind of causal reflexivity.[15] One classic example involves economics education. In two classic studies, Robert Frank, Thomas Gilovich, and Dennis Regan set out to determine whether taking courses in economics has a detrimental effect on members of society by making them more selfish overall (Frank, Gilovich, and Regan, 1993; Frank, Gilovich, and Regan, 1996). To test this theory, the authors recruited a group of college students, some who had previously taken introductory economics courses and others who had not, and they asked the students to play the 'Ultimatum Game'. For those unfamiliar with this game, a pot of money, say ten one dollar bills, is given to 'player one' who decides how the pot should be split between the two players. Then 'player two' decides whether to either accept player one's offer, in which case both take the allotted sums, or to instead reject the offer, in which case both players leave empty handed. According to traditional economic theory, which operates on the assumption that a rational agent will value money in a strictly increasing way, we should predict that player one will tend to make heavily lopsided offers, and player two will tend to accept those heavily lopsided offers. After all, if player one is rational, then she values six dollars more than five dollars, seven dollars more than six dollars, etc., and player two should value one dollar more than zero. So the rational play is for player one to suggest that she takes nine dollars and that player two gets just one. And the rational play is for player two, if given such an offer, to go ahead and accept it.

In fact, heavily lopsided offers have proven extremely rare whenever the Ultimatum Game has been tested on real people. And the few heavily lopsided offers that are given are almost always rejected. With one exception: economics

---

reflexive predictions (Buck 1963; Romanos 1973), performativity (Callon 1998; MacKenzie 2008), and positive or negative reflexive feedback loops (Soros 2013). But what I say below won't depend on any specific analysis of the mechanism behind the phenomenon. All I need, for my purposes, is for predictions in the social sciences to be able to change the beliefs and actions of some of the agents subject to the predictions, and surely that is true for many predictions.

[15] For example, there was an entire recent issue of the *Journal of Economic Methodology* dedicated to it (vol 20 n. 4, 2013).

students. Admittedly, students who have taken some introductory economics courses still rarely offer *extremely* lopsided offers of nine to one, or even eight to two. But on the whole, economics students tend to give more lopsided offers when they are in the position of player one, and they tend to accept more lopsided offers when they are in the position of player two. So it seems that as these students learn about the rational choice model in their introductory economics courses, according to which rational agents value money in a strictly positive way, they tend to use this assumption to predict the behaviour of others when their task is to make the offer. And such students seem to have internalised the normative aspects of the model, which explains why they are more likely to accept lopsided offers. After all, that would be the *rational* thing to do![16] So this is a clear case in which an economic model, and its resulting predictions, has a causal effect on the phenomenon it is intended to model. Knowledge about the model and the predictions it yields changes the beliefs and actions of the very agents subject to the predictions. The model's predictions, at least to some extent, are self-fulfilling.

The sociologist of economics Donald MacKenzie has uncovered a range of historical cases where economic models have exhibited this self-fulfilling character, which he calls 'Barnesian performativity' (after the work of the sociologist and philosopher of science Barry Barnes). One particularly impressive example involves the Black-Scholes (or Black-Scholes-Merton) equation for options pricing (MacKenzie, 2006; MacKenzie, 2008). Prior to 1973, when a famous paper by Fisher Black and Myron Scholes appeared (Black and Scholes, 1973), there was no precise way to put a value on options,[17] and thus the prices of options tended to fluctuate somewhat wildly. The Black-Scholes equation offered a precise method of valuing options, and their market values quickly came to match the prices the formula spat out. As one well-respected economics reference work stated in 1987, 'When judged by its ability to explain the empirical data, option pricing theory is the most successful theory not only in finance, but in all of economics'(Ross et al., 1987: 332; quoted in MacKenzie, 2006: 42). But as MacKenzie argues, the model was itself affecting the values of the options, instead of merely providing an accurate representation of their underlying value. In short, the widespread belief among the traders that the fair values of the options were given by the model, in turn, made the market prices for those options conform to that belief. And when an extreme wave of volatility stuck the markets in the late 1980s, the predictive accuracy of the model quickly crumbled.

From MacKenzie's careful work, we can extrapolate five main characteristics that tend to signal that an economic model is exhibiting a self-fulfilling reflexivity (MacKenzie, 2006: 43-49). First, the model contains idealising assumptions that are strictly speaking false. Second, the model comes with a high

[16] One wonders whether economics education would have the same effects if a non-normative term were used. Thanks to Suzanne Uniacke and Nicholas Southwood on this point.

[17] For those unfamiliar, an option is, roughly speaking, a contract that gives the purchaser the right, but not the obligation, to purchase a certain asset at a later date and at a certain price.

degree of apparent scientific legitimacy.[18] Third, the model is cognitively simple and yet has explanatory depth. Fourth, even those who disagree with the veracity of the model are encouraged to employ it anyway. And, fifth, the actions of agents can make the previously false assumptions on which the model is based more closely approximate the truth. While MacKenzie does not claim that such conditions are jointly sufficient for a model to exhibit causal reflexivity, they greatly increase the likelihood that it will.

Let me quickly walk through how these conditions are met in the case of economics education discussed above. First, the rational choice model holds that agents will value each additional unit of currency in a strictly positive way, which, as the experimental results have shown, is actually false for most people. Second, the rational choice model has a great deal of apparent scientific legitimacy—it is, after all, a core model of the field that arguably has the highest reputation among the social sciences. Third, the rational choice model, while cognitively simple, allows us to explain all kinds of social and economic behaviours, and thus offers a great deal of explanatory depth. Fourth, even a player who denies the assumption of a strictly positive monetary valuation will benefit from using it in a case where the other players all believe it is true. (In other words, if you are player one in an Ultimatum Game, and you are playing against an economics student, it is in your best interest to make more lopsided offers than you normally would, since the person you are paired with is more likely to accept that lopsided offer.) And, fifth, the actions of agents who know about the model can make the model's initially false assumptions come to more closely resemble the truth. This is exactly what happens when the economics students start to accept more lopsided offers—they come to resemble more closely an agent with a strictly positive valuation function for currency.

The case of modelling international greenhouse gas emissions as a game-theoretic tragedy of the commons also meets all of these conditions. First, the tragedy model contains a number of strictly speaking false assumptions. One crucial assumption is that it is in each state's rational self-interest to boost its economic output if possible. This is strictly speaking false because whether something is in a state's rational self-interest ultimately depends upon the state's preferences. The assumption above is only guaranteed to be true if states value economic output in a strictly positive way. To the extent that it even makes sense to talk of a state's preferences, this assumption surely proves false in a wide range of cases, for example, when a state imposes regulations it knows will be overall harmful to the economy. In short, nations seem to care about many things besides merely increasing their economic output. (I'll have more to say on this point in the next section.)

---

[18] For this condition to be met, all that is required is that enough of the relevant actors believe that the model has scientific legitimacy. The model does not actually need to *be* scientifically legitimate, in the sense that it meets some set of minimal criteria for scientific objectivity, empirical confirmation, etc. I thank an anonymous reviewer who caught an ambiguity in my earlier formulation of this condition.

Second, the use of the tragedy of the commons to model the climate problem comes with a great deal of apparent scientific legitimacy. As mentioned earlier, the most recent IPCC report, which is about as close as one gets to an expression of scientific consensus, supports this way of modelling the problem (IPCC, Working Group III, 2014: Ch3).

Third, the model is certainly cognitively simple—it takes but a few paragraphs to get the general idea across to the reader. But, while simple, it offers a seemingly deep explanation of a range of human social problems. In particular, once one grasps the basics of the model, one can very easily make sense of the rather complex failures of our international negotiations. This is truly impressive, given how easy it is to grasp how the tragedy model works.

Fourth, even if the relevant negotiators for a certain nation did not believe the tragedy model to be a correct representation of the international situation, they would still be encouraged to use it to predict the behaviour of the other nations. In a strategic interaction, it not only matters what an actor initially intends to do, but also what the competition believes the actor intends to do. And since the tragedy model is currently the dominant model in use, any outlier nation (i.e., one that actually rejects the model) will be assumed to act in accord with the model's predictions regardless. In such a case, the model will infect the negotiations whether or not one accepts it as the correct representation of reality, and so that outlier nation is encouraged to employ it so it will be able to predict the intentions and actions of the other nations.

And, fifth, the actions of nations can make the initially false idealising assumptions come to more closely approximate the truth. Rather paradoxically, as the tragedy model has increasingly grown in prominence, it seems as though individual nations have come to act more in line with the assumption that they value economic output in a strictly positive way.[19] Each nation's preference for increasing economic output was further confirmed with every failed meeting and every treaty defection. Thus, even if that assumption was once somewhat obviously false, perhaps back around the time of the Rio Earth Summit, it seems to have come to more closely approximate the truth. Nations now really do seem to value economic output in the way the game theorists have assumed they do.[20]

---

[19] Just to give the reader some feel for how the influence of the model has grown, articles citing Hardin's classic article increased by a factor of roughly 12 after 1992 (the year of the Rio Earth Summit), but articles citing it and mentioning the phrase 'climate change' increased by a factor of roughly 81 (6820/84) and the phrase 'greenhouse gas' by a factor of roughly 93 (2140/23) (Google Scholar search as of 5 June 2015). And as the model gained prominence after the Earth Summit, there was also a notable change in the tenor of climate meetings.

[20] An anonymous reviewer has pointed out that there is another possible explanation for why nations have increasingly focused on economic activity (often put in terms of GDP): it is a predictable outcome of their increasing reliance on economic approaches in policy-making more generally. I completely agree that the increasing reliance on economic approaches is also playing a role. But for this fifth condition to be met, all that is required is that it's *possible* for the actions of the relevant agents to change in such a way that the idealizing assumptions more closely

At this point, I should admit that the argument of this section is not intended to be absolutely definitive. As I noted earlier, MacKenzie himself does not claim that these conditions are anything like sufficient conditions. Rather, they are markers that suggest causal reflexivity might be at work. What makes MacKenzie's analyses of cases like Black-Scholes so compelling is the extent of his careful sociological work. To establish, definitively, that the tragedy of the commons model is playing a similar reflexive role in our climate negotiations would require a similar program of careful sociology. In particular, we would have to speak with world leaders and their respective climate negotiators to see whether the model has had an effect on their thinking, and then look for signs that such changes in thought have caused changes in action. I cannot pursue this massive project here. My only intention has been to argue that it is likely that the widespread use of the tragedy of the commons to model climate negotiations is having a causal effect on the negotiations, and, in turn, that we should not necessarily take the model's successful predictions as a reason for pessimism.

4. Dispelling the Tragedy

Suppose that I am right in thinking that the tragedy model is acting as a kind of self-fulfilling prophecy in the case of climate negotiations. What should we do about it? In this section I will discuss some strategies we could employ to try to counteract the model's self-fulfilling tendencies and thus to hopefully dispel the impending tragedy.

The first strategy would be to insist that whenever the tragedy of the commons is presented as a way to model climate negotiations, we should insist that those presenting the model are clear about how the assumptions of the model are not likely to be strictly speaking true. In particular, economists and game theorists involved in the discussion ought to be honest about the widespread nature of altruistic preferences around the world. Although I didn't focus on it earlier, there was a hidden moral in the economics education studies. While students who have taken a few economics courses tend to act in rather selfish ways when put in the Ultimatum Game, the rest of the population really doesn't, since fair splits are actually very common. And this is not just true of American college students,[21] since fair splits are actually common the world over. In one particularly perplexing case, when the Lamalera whalers of Nusa Tenggara in Indonesia played the Ultimatum Game, offers that were *more* than fair were actually very common, and such overly altruistic offers were very often *rejected* (Henrich et al., 2001). The point is that social norms, and perhaps even moral

approximate the truth. This can be true even if there are other causal factors also pushing in the same direction. I thank the reviewer for pressing this concern.

[21] As Henrich, Heine, and Norenzayan (2010) note, we need to be aware that our behavioural studies tend to focus solely on the 'weirdest people in the world'.

values, play an important role in the preferences of real people. We would need some good reason to believe that this could not also be the case for nations.[22]

One might question this suggestion, by arguing that the best explanation of fair offers in the Ultimatum Game is that the player giving the fair offer simply realises that there are certain social norms of fairness, and that if her partner accepts such norms, then she is likely to reject a lopsided offer as punishment. If this is the case, then the player who gives a fair offer might not be acting altruistically at all—she simply realises the likelihood that an unfair offer will be punished and is attempting to maximise her own reward. But this kind of reply doesn't hold up to further scrutiny. For example, there is a close relative of the Ultimatum Game, called the Dictator Game, which has also been widely studied (see Kahneman, Knetsch, and Thaler, 1986; Camerer and Thaler, 1995; Engel, 2011). This game is structurally similar to the Ultimatum Game, except that the second player has no option to 'reject' the offer. Although truly fair offers are less common than in the Ultimatum Game, they are still much more common than the economists would expect. And, surprisingly, fair offers occur even among toddlers as young as 3 years of age (Gummerum et al., 2010).

A second strategy for dispelling the self-fulfilling tragedy would be to take seriously attempts to show that it actually is *not* in the immediate self-interest for states to increase their greenhouse gas emissions, even if we put aside concerns over climate change. Recall that for something to fit the tragedy of the commons model, each individual in the game must gain a benefit by abusing the commons. Take the goat case from earlier, but let us add a bit more realism. Goats cost money, and there will be a point at which the cost of adding an additional goat to the commons will cost more than the financial benefit the farmer gains for adding it. Thus there will be a point where even the farmer who lacks altruistic motives will not want to add another goat. If that new equilibrium point is well short of where the commons gets ruined, a community of rationally self-interested farmers will not ruin the commons. A number of economists have begun to argue that industrialised nations are in a similar situation with greenhouse gas emissions (e.g., Green, 2015; Stern, 2015).[23] For example, nations pay a great cost for burning fossil fuels that is rarely taken into account fully, namely, the heavy costs to health and well-being from air pollution driven ailments like asthma and other respiratory diseases. There are also various other hidden costs. Green (2015) argues that once these costs are all properly tallied, it

[22] Here I do not intend to suggest that there are no reasons to think nations might not act as altruistically as normal humans do. Surely there are. For example, there is some evidence that teams tend to act less altruistically than individuals (Bornstein and Yaniv, 1998; Bornstein, Kugler, and Ziegelmeyer, 2004; Luhan, Kocher, and Sutter, 2007). (Interestingly, gender seems to play a role, and including women in the teams dampens this effect—see Dufwenberg and Muren (2006).) One could also make a reasonable case that it is harder for these preferences to scale up in democratic societies, since democratic leaders need tangible gains to justify re-election. My only point here is that those who present the model rarely, if ever, get into such matters, since they don't typically admit to employing an idealising assumption in the first place.

[23] I thank Holly Lawford-Smith for leading me to this line of research.

turns out that it is actually in the best interest of most of the key nations to *decrease* their greenhouse gas emissions as opposed to maintaining the status quo. Taking such arguments seriously, and encouraging more analyses that might point in that direction, might bring into question whether our climate problem fits the tragedy model, even if we accept the strong assumption that nations will always act solely out of self-interest.

The third strategy for dispelling the self-fulfilling climate tragedy would be to encourage those in the field to examine and engage with alternatives to the tragedy of the commons model, in particular, models that do not entail that tragedy is rather inevitable. For example, some authors have recently shown that international negotiations over greenhouse gas emissions alternatively can be modelled more as a so-called 'bargaining game,' and when modelled as such, the game actually has some solutions (e.g., Finus, 2008; DeCanio and Fremstad, 2013; Smead et al., 2014). As another example, there is a branch of game theory known as 'cooperative game theory,' that has received only a fraction of the attention of its non-cooperative counterpart when it comes to climate change (e.g., Diamantoudi and Sartzetakis, 2006; Diamantoudi and Sartzetakis, 2014).[24] Instead of focusing on the self-interested actions of individual agents in a competitive environment, cooperative game theory analyses how different agents can form mutually beneficial coalitions to maximise the amount of overall value that is created. The idea here is that different agents can coordinate their actions in certain ways that will maximise the collective benefit, and the value thus created can be split amongst the respective coalitions.[25] When this framework is used to model climate change, the problem looks solvable, increasingly so when smaller coalitions are formed first (Cole, 2015). The last alternate model I will mention has been called a 'collective risk social dilemma', where groups play a public goods game in which they must contribute resources in order to avert a potentially catastrophic loss (e.g., Milinski et al., 2008; Tavoni et al., 2011; Vasconcelos, Santos, and Pacheco, 2013). When placed in these kinds of dilemmas, real people in lab experiments are sometimes able to avert catastrophe (see Milinski et al., 2008).[26]

I admit that these alternative models contain their own idealising assumptions, and so there may not be an obvious theoretical reason to prefer a cooperative game analysis, bargaining game analysis, or collective risk analysis of the situation over the tragedy of the commons model. Perhaps worse, we might actually have some epistemic reasons to favour the tragedy model, since it carries a fair bit of empirical confirmation. After all, international climate negotiations

[24] For more general information on cooperative game theory, see Moulin (2014). I thank Effrosyni Diamantoudi for leading me to this line of research, and I thank Daniel Hausman for additional leads.

[25] I think there may prove to be some interesting parallels between these cooperative models and other less formal accounts that attempt to solve collective action problems by invoking social norms such as integrity (e.g., Hourdequin, 2010; Hourdequin, 2011).

[26] I thank Justin Bruner for leading me to this line of research.

have generally gone very badly over the past couple of decades, and any model that gives an optimistic spin on the situation seems to fly in the face of this reality. Yet this is why self-fulfilling models can put the scientist into such a perplexing position. If the tragedy model is self-fulfilling in our current international context, then it will seem to a scientist that she has good reason to have confidence in the model because of all predictive accuracy the model has shown. And the community's continued confidence in the model, which these confirmations help to bolster, can then serve to increase the model's hold on reality. It is only when scientists and policy makers come to reject the well-confirmed model that its hold can begin to slip. In the case of the climate tragedy, we have good reason to ensure it does.

5. Conclusion

This article aims to give the reader some reasons to resist the kind of pessimism about our ability to solve the climate crisis that seems rationally forced upon us, given how climate negotiations have tended to play out. I have argued that the dominant game theoretic analysis of climate negotiations, the tragedy of the commons, is likely a kind of self-fulfilling prophecy. In particular, the tragedy model fits the characteristics of other self-fulfilling models in the social and behavioural sciences. If this is true, then perhaps the previous accurate predictions of the model really should not lead us to pessimism. Some very recent events, like the successful negotiations between China and the US last year and the overall success of the recent Paris meeting ('COP 21'), might suggest the self-fulfilling tragedy is starting to lose its hold. I sketched some strategies that we could use to dispel the self-fulfilling tragedy, including fully acknowledging that humans tend to have altruistic preferences, further examining whether increased greenhouse gas emissions come at a net cost as opposed to a net benefit, and examining whether other game-theoretic models, those that lack the tragic conclusion, might be preferable ways to model the negotiations. I hope this leaves room for optimism.

Acknowledgements

I would like to thank Matthew Barker, Justin Bruner, Bert Bumgartner, Steve Clarke, Ray Dacey, Effrosyni Diamantoudi, Joshua Filler, Donald MacKenzie, Trevor Pearce, Amanda Szabo, Nicholas Southwood, and Suzanne Uniacke for helpful comments, suggestions, or discussion. I would also like to thank the participants at the Inland Northwest Philosophy Conference and CAPPE's Works in Progress Group, as well as the audiences at Concordia University in Montreal, St. Mary's College of Maryland, La Trobe University, and Australian National University. I would like to thank Tamara Browne for her astute editing. Finally, I would like to thank an anonymous referee whose challenges and suggestions helped me greatly improve the paper.

References

Barrett, S. 2003. *Environment and Statecraft : The Strategy of Environmental Treaty-Making*. Oxford: Oxford University Press.

Bernauer, T. 2013. 'Climate Change Politics.' *Annual Review of Political Science* **16**: 421–448.

Binmore, K. 2007. *Game Theory: A Very Short Introduction*. Oxford: Oxford University Press.

Binmore, K. 2012. *Playing for Real Coursepack Edition: A Text on Game Theory*. Oxford: Oxford University Press.

Black, F., and M. Scholes. 1973. 'The Pricing of Options and Corporate Liabilities.' *Journal of Political Economy* **81**: 637–654.

Bornstein, G., T. Kugler, and A. Ziegelmeyer. 2004. 'Individual and Group Decisions in the Centipede Game: Are Groups More 'rational' Players?' *Journal of Experimental Social Psychology* **40**: 599–605.

Bornstein, G., and I. Yaniv. 1998. 'Individual and Group Behavior in the Ultimatum Game: Are Groups More 'rational' Players?' *Experimental Economics* **1**: 101–108.

Buck, R. 1963. 'Reflexive Predictions.' *Philosophy of Science* **30**: 359–369.

Callon, M. 1998. *The Laws of the Markets*. Oxford: Blackwell.

Camerer, C., and R. Thaler. 1995. 'Anomalies: Ultimatums, Dictators and Manners.' *The Journal of Economic Perspectives* **9**: 209–219.

Cole, D. 2015. 'Advantages of a Polycentric Approach to Climate Change Policy.' *Nature Climate Change* **5**: 114–118.

DeCanio, S., and A. Fremstad. 2013. 'Game Theory and Climate Diplomacy.' *Ecological Economics* **85**: 177–187.

Diamantoudi, E., and E. Sartzetakis. 2006. 'Stable International Environmental Agreements: An Analytical Approach.' *Journal of Public Economic Theory* **8**: 247–63.

Diamantoudi, E., and E. Sartzetakis. 2014. 'International Environmental Agreements: Coordinated Action under Foresight.' *Economic Theory* **59**: 527–46.

Dietz, T., E. Ostrom, and P. Stern. 2003. 'The Struggle to Govern the Commons.' *Science* **302**: 1907–1912.

Dodds, W. 2005. 'The Commons, Game Theory and Aspects of Human Nature that May Allow Conservation of Global Resources.' *Environmental Values* **14**: 411-25.

Dufwenberg, M., and A. Muren. 2006. 'Gender Composition in Teams.' J*ournal of Economic Behavior & Organization* **61**: 50–54.

Engel, C. 2011. 'Dictator Games: A Meta Study.' *Experimental Economics* **14**: 583–610.

Ferraro, F., J. Pfeffer, and R. Sutton. 2005. 'Economics Language and Assumptions: How Theories Can Become Self-Fulfilling.' *Academy of Management Review* **30**: 8–24.

Ferraro, F., J. Pfeffer, and R. Sutton. 2009. 'How and Why Theories Matter: A Comment on Felin and Foss (2009).' *Organization Science* **20**: 669–675.

Finus, M. 2008. 'Game Theoretic Research on the Design of International Environmental Agreements: Insights, Critical Remarks, and Future Challenges.' *International Review of Environmental and Resource Economics* **2**: 29–67.
Frank, R., T. Gilovich, and D. Regan. 1996. 'Do Economists Make Bad Citizens?' *Journal of Economic Perspectives* **10**: 187–92.
Frank, R., T. Gilovich, and D. Regan. 1993. 'Does Studying Economics Inhibit Cooperation?' *Journal of Economic Perspectives* **7**: 159–171.
Gardiner, S. 2006. 'A Perfect Moral Storm: Climate Change, Intergenerational Ethics and the Problem of Moral Corruption.' *Environmental Values* **15**: 397-413.
Gardiner, S. 2011. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change.* Oxford: Oxford University Press.
Green, F. 2015. 'Nationally Self-Interested Climate Change Mitigation: A Unified Conceptual Framework.' *London School of Economics and Political Science, CCCEP Working Paper*, no. 224.
Gummerum, M., Y. Hanoch, M. Keller, K. Parsons, and A. Hummel. 2010. 'Preschoolers' Allocations in the Dictator Game: The Role of Moral Emotions.' *Journal of Economic Psychology* **31**: 25–34.
Hardin, G. 1968. 'The Tragedy of the Commons.' *Science* **162**: 1243–1248.
Harris, P. 2010. *World Ethics and Climate Change: From International to Global Justice.* Edinburgh: Edinburgh University Press.
Henrich, J.,et al. 2001. 'In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies.' *The American Economic Review* **91**: 73–78.
Henrich, J., S. Heine, and A. Norenzayan. 2010. 'The Weirdest People in the World?' *Behavioral and Brain Sciences* **33**: 61–83.
Hourdequin, M. 2010. 'Climate, Collective Action and Individual Ethical Obligations.' *Environmental Values* **19**: 443-464.
Hourdequin, M. 2011. 'Climate Change and Individual Responsibility: A Reply to Johnson.' *Environmental Values* **20**: 157-162.
International Panel on Climate Change. 2014. *Fifth Assessement Report (AR5). Working Group III, Chapter 3. Social, Economic and Ethical Concepts and Methods.* http://www.ipcc.ch/pdf/assessment-report/ar5/wg3/ipcc_wg3_ar5_chapter3.pdf (accessed 7 July 2016).
Johnson, B. 2003. 'Ethical Obligations in a Tragedy of the Commons.' *Environmental Values* **12**: 271-81.
Johnson, B. 2011. 'The Possibility of a Joint Communiqué: My Response to Hourdequin.' *Environmental Values* **20**: 147-156.
Kahneman, D., J. Knetsch, and R. Thaler. 1986. 'Fairness and the Assumptions of Economics.' *The Journal of Business* **59**: S285–300.
Kopec, M. 2011. 'A More Fulfilling (and Frustrating) Take on Reflexive Predictions.' *Philosophy of Science* **78**: 1249-1259.
Lloyd, W. 1833. *Two Lectures on the Checks to Population.* Oxford: Collingwood.
Luhan, W., M. Kocher, and M. Sutter. 2007. 'Group Polarization in the Team Dictator Game Reconsidered.' *Experimental Economics* **12**: 26–41.

MacKenzie, D. 2006. 'Is Economics Performative? Option Theory and the Construction of Derivatives Markets.' *Journal of the History of Economic Thought* **28**: 29–55.
MacKenzie, D. 2008. *An Engine, Not a Camera: How Financial Models Shape Markets.* Cambridge, Massachusetts: MIT Press.
Merton, R. 1948. 'The Self-Fulfilling Prophecy.' *The Antioch Review* **8**: 193–210.
Milinski, M., et al. 2008. 'The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change.' *Proceedings of the National Academy of Sciences* **105**: 2291–2294.
Moulin, H. 2014. *Cooperative Microeconomics: A Game-Theoretic Introduction.* Princeton, New Jersey: Princeton University Press.
Ostrom, E. 2015. *Governing the Commons.* Cambridge: Cambridge University Press.
Ostrom, Elinor, et al. 1999. 'Revisiting the Commons: Local Lessons, Global Challenges.' *Science* **284**: 278–282.
Romanos, G. 1973. 'Reflexive Predictions.' *Philosophy of Science* **40**: 97–109.
Rosenthal, R., and L. Jacobson. 1968. *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development.* Holt, Rinehart & Winston.
Ross, S., et al. 1987. *The New Palgrave Dictionary of Economics.* London: Macmillan.
Sandler, T. 2004. *Global Collective Action.* Cambridge: Cambridge University Press.
Simon, H. 1954. 'Bandwagon and Underdog Effects and the Possibility of Election Predictions.' *Public Opinion Quarterly* **18**: 245–253.
Smead, R., R. Sandler, P. Forber, and J. Basl. 2014. 'A Bargaining Game Analysis of International Climate Negotiations.' *Nature Climate Change* **4**: 442–445.
Soroos, M. 1997. *The Endangered Atmosphere: Preserving a Global Commons.* Columbia: University of South Carolina Press.
Soros, G. 2003. *The Alchemy of Finance.* Hoboken: John Wiley & Sons.
Soros, G. 2013. 'Fallibility, Reflexivity, and the Human Uncertainty Principle.' *Journal of Economic Methodology* **20**: 309–329.
Stern, N. 2015. 'Economic Development, Climate and Values: Making Policy.' *Proceedings of the Royal Society B* **282**: 20150820.
Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel. 2011. 'Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game.' *Proceedings of the National Academy of Sciences* **108**: 11825–11829.
Vasconcelos, V., F. Santos, and J. Pacheco. 2013. 'A Bottom-up Institutional Approach to Cooperative Governance of Risky Commons.' *Nature Climate Change* **3**: 797–801.