Theory Roulette: Choosing that Climate Change is not a Tragedy of the Commons

JAKOB ORTMANN

Department of Philosophy, University of Bayreuth
Email: jakobortmann@gmail.com
https://orcid.org/0000-0002-3967-1333

WALTER VEIT

School of History and Philosophy of Science, University of Sydney
and
The Centre for Philosophy of Natural and Social Science,
London School of Economics and Political Science
Email: wrwveit@gmail.com
https://orcid.org/0000-0001-7701-8995

ABSTRACT

Climate change mitigation has become a paradigm case both for externalities in general and for the game-theoretic model of the Tragedy of the Commons (ToC) in particular. This situation is worrying, as we have reasons to suspect that some models in the social sciences are apt to be performative to the extent that they can become self-fulfilling prophecies. Framing climate change mitigation as a hardly solvable coordination problem may force us into a worse situation, by changing real-world behaviour to fit our model, rather than the other way around. But while this problem of the performativity of the ToC has been noted in a recent paper in this journal by Matthew Kopec, his proposed strategies for dealing with their self-fulfilling nature fall short of providing an adequate solution. Instead of relying on the idea that modelling assumptions are always strictly speaking false, this paper shows that the problem may be better framed as a problem of underdetermination between competing explanations. Our goal here is to provide a framework for choosing between this set of competing models that allows us to avoid a 'Russian Roulette'-like situation in which we gamble with existential risk.

KEYWORDS

Tragedy of the Commons, climate change, philosophy of science, game theory, ethics of economics

Environmental Values
Submitted 26 January 2021; accepted 7 January 2022; fast track 7 March 2022

 doi: 10.3197/096327122X16452897197784

1. FRAMING THE PROBLEM

Climate change (CC) is often modelled as a Tragedy of the Commons (ToC). Indeed, this has happened so many times that it appears to have evolved into a paradigm example for game theory, microeconomics and political science as the ultimate Tragedy of the Commons: the Prisoner's Dilemma (PD) of doom.[1] The general idea behind the ToC was elegantly summarised in the seminal paper by Garret Hardin:

> 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. […] 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. Freedom in a commons brings ruin to all. (Hardin 1968: 1244)

Importantly, there seems to be little doubt among the scientific community that this is the right way to model the problem of climate change mitigation. The Intergovernmental Panel on Climate Change (IPCC), for instance, ascribes 'high confidence' (IPCC 2014: 211) to the suspicion that climate change is indeed correctly characterised as a version of this well-known game-theoretic story of cattle and herdsmen.

The implications of this model, however, are not just worrisome, they are frightening. For if it is correct to model climate change as a ToC, then there is little room left for optimism that our political means will be apt to tackle the challenges of climate change mitigation. This concern mainly arises because historic methods for dissolving the ToC (i.e. privatisation and top-down regulation) do not appear to help much in the specific case of regulating greenhouse gas (GHG) emissions, as the damage caused by emissions is dispersed, both in time and geography. These characteristics gave rise to a class of worries that we will refer to as deterministic pessimism. Matthew Kopec, who has raised this concern forcefully, was thus inclined to frame our current situation as the climate crisis being seemingly 'rationally forced upon us' (Kopec 2017: 15). In a similar vein, Hardin framed the tragedy in terms of a lack of a 'technical solution' (Hardin 1968: 1248), meaning that the pasture could not be expanded or replaced. The deterministic pessimism radiating from these characterisations suggests that if our situation is sufficiently close to the ToC, climate change is bound to appear as an unsolvable problem. This alone provides plenty of reason to put the ToC under much more rigorous scrutiny, especially since the confidence in it seems so incontrovertible.

1. For a proper disambiguation of the terms *Prisoner's Dilemma (PD)* and *Tragedy of the Commons (ToC)* in the case of climate change see e.g. MacLean (2015). In this paper, however, we will treat the ToC as a variation of a PD with incremental decisions by more than one agent. Related terms are *commons problem, common pool resource problem* and *externalities problem.*

Even more worrying, however, is the fact that there is now plenty of empirical evidence that at least some models in the social sciences are 'performative' to the extent that they are apt to become self-fulfilling prophecies. When used and implemented as models in scientific or political discourse these models have a propensity to interfere causally with what they merely want to describe (see Mackenzie 2006, 2008).[2] The characteristics of such self-fulfilling performative models are roughly identified as (1) containing idealising assumptions that are strictly speaking false, (2) obtaining a high degree of scientific legitimacy and (3) being cognitively simple while having a significant explanatory depth (see Mackenzie 2006: 43–46 and Kopec 2017: 9–10).[3] According to Kopec, the ToC model applied to climate change is highly likely to satisfy these characteristics and thus is likely to be a self-fulfilling prophecy. This problem is the main concern of this paper.

If Kopec's claim is true, asking the question of how confident we are about the ToC is not only a purely positive-descriptive endeavour, but one with severe and dangerous consequences of ethical dimension. This is because when a self-fulfilling behavioural model creeps into the minds of decision-makers, it kicks off a positive feedback loop that eventually generates the very evidence everybody counts on to check for model-world-fit. With such a loop running in the background we can never be sure whether the currently observed failure to mitigate climate change stems from the causal relationships established by the model – or mainly because people believe that this is how the world works and act accordingly. Therefore, it would be a scientific and ethical mistake to base our model choice solely on model-world-fit without checking on the effects that these choices can cause.

In an attempt to maintain at least some optimism about humanity's ability to alleviate the climate crisis as well as to avert the potential self-fulfilling performativity of the commons model, Kopec suggests some strategies to counter its potentially damning performative effects. One of them is to insist in every presentation of the ToC model for climate change negotiations that 'the assumptions of the model are not likely to be strictly speaking true' (Kopec 2017: 12).

While we agree with Kopec's general point about the danger of self-fulfilling performativity in the case of the ToC, we suggest that the solution strategy he proposes needs to be substantially improved. As we will show in section 2, all models contain 'falsehoods' and idealisations. Models are deliberately designed only to represent a part of the world and shed light on some conditional causal connections – that is, if they succeed. As a result, talk of the strictly

2. Note that, in the literature, the term *performativity* can denote both self-fulfilling and self-refuting effects. In the case of the ToC, however, we are only interested in the former variant. A related and sometimes interchangeably used term is *reflexivity*.

3. These properties should not be seen as anything like a definition in terms of necessary and sufficient conditions for performativity to obtain, but rather as factors that correlate with and exacerbate performativity as a problem for policy-makers.

speaking false assumptions of the ToC remains trivial and is thus unlikely to be helpful.

In contrast, this paper aims to put forward a more robust argumentative foundation to deal with the same problem, beginning in section 3. We argue that the recognition of the general underdetermination of models by the empirical evidence should lead us to endorse a more pluralistic modelling approach that recognises modelling as an activity with myriad goals and constraints – not just as an attempt to fit with the real world.

Because of the underdetermined palette of modelling options, we are in need of a pragmatic framework for model choice that pays attention to their context, and, perhaps more importantly, some guidance on how to communicate them effectively to the public. One suggestion for a requirement of such a framework on communication emerges from a particular reading of underdetermination that we will spell out in a subsequent section: the combination of dangerous self-fulfilling performativity together with underdetermination of behavioural theories renders us part of a perverse kind of Russian Roulette that we shall dub Theory Roulette.

2. ON STRICTLY SPEAKING FALSE ASSUMPTIONS

At first sight, Kopec's strategy for dealing with the performativity of climate change mitigation models appears reasonable, emphasising that assumptions are 'strictly speaking false'. Since the ToC is a model, it necessarily *must* employ idealised assumptions specifically to enable it to say anything useful, so it is true that the ToC model by extension is strictly speaking false (see e.g. Weisberg 2015). Furthermore, it is certainly correct to point out the boundaries and limitations of our models, even more so when we are worried about potentially dangerous performativity. Kopec's strategy is also in line with a tradition of criticising economic models like the ToC for being too simple, unrealistic and ignoring important features of the real world. In light of contemporary work in the philosophy of science, however, Kopec's strategy turns out to be too weak for comfort.

Mainly, this is because it appears particularly unhelpful to add the clarification that the ToC is strictly speaking false, for it suggests that there is something especially problematic that is not shared by other models. But as both scientists and philosophers have long argued, *every* assumption and *every* conclusion of *every* theory that we will *ever* come up with may very well be considered as not being likely to be true. The necessity for idealisation renders all models 'false' in this strict sense and it must therefore be rather considered a *feature, not a bug* of a model. Imagine, for example, standing in front of a subway map – a thought experiment often used in the philosophical literature on models (Kitcher 2001; Weisberg 2015). With the explicit goal to travel from

point A to point B, you require a map/model that helps you in doing exactly that and not a map that resembles the real world as closely as possible. The common abstractions of subway maps render them less approximative to the real world than more detailed maps. Nevertheless, they provide you with relevant and useful insights for your specific task by not obscuring these insights with information that is irrelevant to you. Models or theories that include *all* the variables would simply be unusable. Variants of arguments along this line of false models still being useful and explanatory have been defended in the past, for example by Uskali Mäki in his account of models as isolations (Mäki 2009; for an overview of other arguments see Weisberg 2015).

Furthermore, Kopec's suggestion of framing the ToC as a strictly speaking false model appears to urge us in the direction that we need to add substantially more empirical data and track the complexity of the world more closely within our model. But that may very well not be useful for making general policy decisions if what we are engaged in is providing useful 'maps'. Sometimes, general and simple models are precisely what we need in a political context because the problem at hand is otherwise made unwieldy. Even among decision-makers with little familiarity with the methodological discussions in the philosophy of science, it is uncharitable to assume that politicians will simply take these models as revealing anything like an *a priori* truth. Therefore, merely pointing out that the ToC relies on strictly speaking false assumptions is hardly news to anybody. While we agree with many of Kopec's actual recommendations, we fear that this core of his argument will hardly be able to convince others that the ToC is, in fact, unhelpful and fails to shed light on some relevant mechanism that leads to free-riding. Therefore, a proper strategy for alleviating a self-fulfilling climate change tragedy cannot rely on the mere fact of idealisation. Instead, we argue, it must engage with the performativity of model-choices directly.

One argument we put forward to deal with the potentially ill-placed confidence in the ToC model arises from a particular reading of the Quinean *argument of underdetermination* (Quine 1951). According to Quine, the empirical data available is at any time insufficient to decide reliably between co-existing theories that are (1) compatible with a given finite set of observations and (2) mutually contradictory (see Stanford 2017). Intuitively, this can be understood analogously to the way a system of mathematical equations can be underdetermined, for example a system with two equations and three variables: an equation system that is underdetermined like this has either none or an infinite number of possible solutions. Similarly, for any theory that explains a finite set of data we are licensed to assume that there may exist alternative theories that we cannot rule out with the help of the given data alone. Quine, who has endorsed such a strong underdetermination thesis, concluded that 'the considerations which guide [someone] in warping his scientific heritage to fit

his continuing sensory promptings are […] *pragmatic*'. (Quine 1951: 43; emphasis added).

This, as we will argue, directly relates to the discussion of whether the ToC is a fitting model for climate change mitigation. In particular, the argument of underdetermination contributes to the problem at hand in two ways. First, it provides further reason why a perfect model-fit to the real world should not be the *non plus ultra* in model choice, since the very question of which model has a better fit to the target phenomenon is empirically underdetermined. Secondly, it provides a first glimpse on how to cope with that prevalence of falsehood and error in theories: a form of pragmatism. We will later return to this important second point.

Moreover, underdetermination has implications for what we may expect from the practice of modelling in general. All too often do the public and policy-makers alike share the unfortunate view that science is in the game of providing something like the one, true model, with old models continuously being discarded and replaced by more general, precise and accurate ones. The problem with this view, however, is that there cannot be a single model that could capture the complexity of a phenomenon like climate change mitigation. If we want a model to be as general as possible, there will necessarily be trade-offs with other epistemic values such as predictive power, precision and causal detail (see Levins 1966). For this reason, it is important to emphasise that when we are dealing with complex sciences such as, for example, ecology or weather forecasting, we rely upon a set of multiple models to accommodate the various trade-offs between different epistemic desiderata. We use multiple models precisely because all models are strictly speaking false in the sense in which Kopec describes the ToC model. We should therefore not reconceive this as an attempt to find the one true model. This way of thinking is tempting, but it may well be misleading. To develop a compelling theory is to create a plethora of models that stand to each other in various robust and mutually illuminating relationships, which is precisely how we deal with the underdetermination of theories (see Veit 2021). Appreciating underdetermination provides us with a useful safeguard against the ubiquitous and potentially dangerous confidence that climate change is obviously a ToC and helps us to address the problem of performativity more effectively.

First, in section 3, we show that the case of climate change mitigation similarly suffers from a problem of underdetermination, with the ToC being only one model among many others. Second, this will force us to make explicit the pragmatic criteria for choosing between these co-existing attempts at explanation. In section 4, we propose a pragmatic framework to help us do exactly that, which we dub Theory Roulette as both a useful metaphor and a literal description of a dangerous public game with which we are faced. Together, both parts suggest a strategy for alleviating a self-fulfilling tragedy; a strategy that aims to go beyond a trivial emphasis on strictly speaking false assumptions.

3. UNDER(DETER)MINING THE TRAGEDY OF THE COMMONS

How does underdetermination play out specifically for the ToC when employed as a descriptive model for current and future mitigation failure? At a basic level, we can derive from the argument of underdetermination that multiple different explanations can exist for the currently observed lack of mitigation. This in itself does not say anything about the plausibility of these alternatives, but at least motivates one to look out for them when concerned about the insolvability of climate change as a ToC.

For example, a minimal extension of the standard ToC is to explicitly assign some form of prosocial preference to the assumed agents. One might be justified in believing that these agents are now more realistic (i.e. they resemble real humans more closely), as prosocial preferences may be considered to be revealed by the factual presence of altruistic behaviour. These prosocial preferences do not change the general assumption that ToC agents are selfishly rational. However, it may be argued that thereby the general payoff matrix has changed such that the pull towards mutual demise is less strong. When prosocial preferences are present, the difficulty of solving collective action problems, like climate change mitigation, is generally thought to be substantially reduced (see e.g. Kline et al. 2018; Ackermann and Murphy 2019; Tilman, Dixit and Levin 2019).

We are now prompted with a set of two game-theoretic set-ups modelling the same phenomenon. One is more pessimistic than the other, but both models are generally flexible enough that it could arguably be maintained for each that they are consistent with the historically observed failure of climate change mitigation; that is, they are consistent with the evidence available. Which of these two models has a better fit with the real-world problem? We cannot answer this by looking at the finite set of evidence alone. The situation is empirically underdetermined.

To deal with the obstacles underdetermination poses for model selection, philosophers of science have proposed a variety of approaches. Some have emphasised certain truth-conducive values that should guarantee a sufficient degree of objectivity in the face of underdetermination, such as aspects that scientists value in their theories, e.g. simplicity, coherence, plausibility and other epistemic values or theoretic virtues (see Kuhn 1962; Longino 1996; Douglas 2003). Furthermore, under the condition that performativity is a real threat, our model choice may not solely be intended to fit the real-world target, but also – and much more importantly – help us to avoid a climate change catastrophe. Therefore, as we will argue later, the performativity of our model choice can also warrant the use of non-epistemic values such as the aversion of risk or, for that matter, the worst-case scenario.

Another relatively recent approach for dealing with underdetermination is to proactively embrace model pluralism as an outcome and necessity of

underdetermination, rather than trying to resist it. Model pluralism recognises model diversity as a strength rather than a weakness that must be eliminated, by recognising that complex phenomena have different aspects that require multiple and different models for explanations that serve different purposes (see Veit 2020). So, without even waiting for *ex-post* (in)validation, we can *ex-ante* assume that a single model like the ToC will not suffice as a descriptive behavioural model after all. We can expect that multiple models for multiple aspects are needed, simply by virtue of our epistemic uncertainty and the performativity of the situation.

The most widespread usage of model pluralism in practice comes in the form of robustness analysis, which introduces small perturbations into our base model (just as we have done in the prosocial preferences example above), thereby creating a whole set of different models that are subsequently compared against each other. Because it is hard to test an individual model, we instead rely on a large family of models with varying assumptions to gain confidence in the robustness of the processes in the model, even in the absence of real-world corroboration (see Aydinonat 2018). Importantly, the result of this procedure is not to discard 'lower performing' or 'more inaccurate' variations out of hand. Rather, it is a continuously ongoing procedure (with no definite end result) of actively trying to find models that jointly illuminate highly complex real-world relations instead of trying to find one general model that captures everything, with modelling being more like a craft than a rule-based procedure (see Cartwright 2019 and Veit 2021).

Accordingly, instead of potentially over-committing to the ToC when it is likely to be self-fulfillingly performative, we suggest taking seriously the implications of model pluralism and treating it as just one model among many. In the following sections, we will provide a set of exemplary models to show that there are plausible alternatives that do not force us into the pessimistic confines of a ToC.

An answer to the question of which model in the exemplary set of classical ToC and ToC with prosocial preferences better fits climate change mitigation, of course, does not have to rely on theoretical considerations on grounds of underdetermination only. Another approach would be to remark that both variations are simply too vague and imprecise in their detail to be useful for making empirical predictions and policy evaluations. This problem of vagueness, for instance, becomes apparent by the simple fact that advocates of climate change being a ToC are often unclear about whether the considered agents are supposed to resemble individuals, nations, national leaders, or all of them at the same time (e.g. IPCC 2014: 211). Arguably, deciding on whoever the agents are supposed to resemble is a quintessential assumption that plays a big role in whether we think a specific payoff matrix is a good real-world fit or not. And once the different implications of these assumptions have been worked out, it arguably seems that we are hard-pressed to present more than one model and

not to treat the national leader equally to the commoner – i.e. as sheep farmers on a tight pasture.

Turning from underdetermination to vagueness like this manifests a supplementary step of this proposed strategy: beyond suggesting variations and alternative explanations (that are likewise underdetermined when being tested against observed behaviour), it might be worth attacking the ToC directly. That is, to figure out where exactly the boundaries of the explanatory power of the ToC lie and in which ways empirical evidence has failed to support the very implications that have motivated its deterministic pessimism to begin with. This would mean showing that the ToC is a short-cut heuristic at best, little more than an imprecise story that does not suffice as a descriptive model of climate change mitigation. We will refer to this as attempting a *non-trivial* falsification as opposed to relying on strictly speaking false assumptions. This attempt, we think, captures the main idea of Kopec's proposed strategy better and it may turn out in two different ways, as is illustrated in Figure 1.

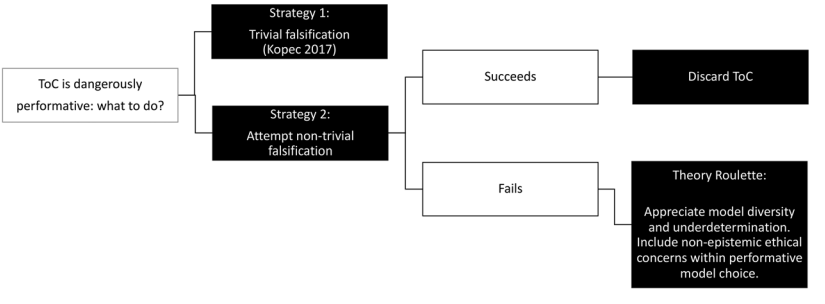

Figure 1. Decision-tree for dealing with the self-fulfilling nature of the ToC. Source: author's illustration.

First, if an attempt at non-trivial falsification succeeds, Kopec's strategy would have to be bolstered up, in that the ToC model for climate change is not only strictly speaking false, but also *plainly* speaking. The ToC would have to be abandoned due to a bad fit with the real world and one would replace it with more accurate models, with potentially more optimistic predictions. And if they are optimistic, self-fulfilling performativity would no longer constitute a necessarily unwelcome problem. In section 3.1, we will roughly sketch out various promising attempts that aim to achieve exactly that.

Second, if such a 'falsification' does not succeed – and this is our main motivation for this paper – then the ToC is still among the underdetermined candidates for a descriptive model of climate change mitigation failure. Mapping out this selection of some alternative candidates will be the task of section 3.2. To respond adequately to this problem, we have to recognise that performativity forces us into the midst of playing an unwelcome 'Theory Roulette' and

that we should rely on the right pragmatic criteria to avoid the worst-case scenario in which an almost 'random' model choice will force us into the confines of an actual ToC, which was the very thing we were trying to avoid.

3.1. Attempting non-trivial falsification

Since we have already seen that employing strictly speaking false assumptions is not an adequate property for a model to be considered useless, an attempt at non-trivial falsification is going to be a more intricate endeavour. It is intricate because the outcome of this attempt can no longer be a simple binary answer, true or false. Instead, it will have to be an answer of degree that, in turn, is highly dependent on aspects like which phenomenon exactly under which conditions is the subject of a particular ToC portrayal.[4]

As a start, this would include disambiguating the earlier mentioned vagueness about who, precisely, the agents are supposed to mirror. Are they representatives of nations that failed to reach and enforce adequate agreements in Kyoto, Copenhagen, Paris and Glasgow? Are they private households that seek to minimise their expenditure on power consumption? Or are they parents who prefer to use their air-conditioned car to drive their children to primary school because it seems more convenient and safer than using a bicycle? In all these cases it has to be asked whether the ToC is the best available model, and the answer may well vary from one case to another.

One way of disentangling this collection of potential ToC instantiations was suggested by Elinor Ostrom over a decade ago, using what she calls a polycentric approach (see Ostrom 2009). Her critical review of the ToC being used to model climate change mitigation failure is based on two grounds: the first is 'the existence of multiple externalities at small, medium, and large scales within the global externality' (Ostrom 2009: 9). This directly corresponds to the earlier identified vagueness about whom the agents are supposed to resemble, that is, choosing the appropriate level of scale and deciding on whether the agents are supposed to resemble (for example) people, nations or national leaders. According to Ostrom, it is not a sound scientific approach only to look at one particular scale for the costs and benefits of GHG mitigation, but instead the multiplicity of effects of diverse actions on multiple scales and their reciprocating influence must be taken into account (Ibid.: 32–35). By ignoring such multi-scale complexity one may legitimately argue that the ToC is too simplistic to be adequate for our purpose of modelling the problem of climate change – similar to how a subway map that does not show you all the available subway lines is exceeding a critical degree of idealisation for the specific purpose for which it is required.

4. This idea, for example, has also been summarised in the adequacy-for-purpose view of models (see Parker 2020) in which the fit of the model to the real world is evaluated in terms of the reasons why we want to rely on the model.

Furthermore, Ostrom offers compelling criticism of the blatant lack of empirical evidence for the conventional ToC predictions. The unambiguous (and in the case of climate change, frightening) predictions of zero cooperation, as the ToC and deterministic pessimism suggest, are simply not supported by observation: 'While many instances of free-riding are observed in the array of empirical research, a surprisingly large number of individuals facing collective action problems do cooperate' (Ibid.: 10). This insight cannot be overstated for a model of which the supposed paradigm case is the largest potential humanitarian crisis in history. As well as providing a book-length analysis of these empirical findings, Poteete, Ostrom and Janssen (2010) call for an updated theory of collective action that accounts for diverse organising of commons at multiple levels. This 'encourages experimental efforts at multiple levels, as well as the development of methods for assessing the benefits and costs […] in one type of ecosystem and comparing these with results obtained in other ecosystems' (Ostrom 2009: 39).

Another approach is to question the assumption of whether climate change mitigation even meets the criteria for being a common pool resource. Various concerns about this crucial assumption, which is often taken for granted, have been raised, for instance, by Anthony Patt (2017). One of these concerns is that there do indeed exist potential technical solutions yielding trivial, if not negative medium-term costs of eliminating GHG emissions (see Edenhofer, Bauer and Kriegler 2005; Patt 2017: 2), which goes against the ToC requirements stipulated by Hardin (1968) that a technical solution to the commons problem must not exist for the tragedy to occur.

Whether technical solutions exist is important for the question of whether climate change mitigation meets the criteria of a common pool resource. Consider that in the framing of the farmer–pasture story, having a technical solution would mean that farmers suddenly lose interest in the pasture because they have developed better alternative technologies that do not require the common pasture (in other words, the use of the commons is not a source of rivalry anymore). And if there is no interest in the commons then there is no Tragedy of the Commons. According to Patt, this important insight of trivial costs is mainly driven by the field of evolutionary economics, with the observation that, for example, 'policies to expand renewable energy also make them cheaper' (Patt 2017: 2), which is highly relevant in a context where fossil fuel sources of energy are heavily subsidised.[5] And indeed, for example, by now the levelised cost of electricity (LCOE) for new renewable energy source power plants in many cases has already sunk beneath the LCOE of new fossil fuel plants (Capros et al. 2016). Thus, there is substantial hope that market dynamics like these change the payoff matrix sufficiently, so that GHG emissions do not constitute a proper common pool resource. Not only is the alignment of preferences of the ToC possibly wrong, but also its spirit of a lack of technical

5. See also Naam 2013.

solutions. A naïve endorsement of the ToC model may not only have stifled research into technical solutions, but it may also have blinded us to the rapid progress that has been made.

A third and final approach to a non-trivially falsifying ToC we want to highlight is from Northcott and Alexandrova (2015): a straight-up refutation of the idea that a simple game-theoretic model like the Prisoner's Dilemma (PD) can be causally explanatory. While we think their conclusion is arguably too strong, they are right to warn of a conflation between the story we tell with a model and the model itself. Consider that, also for climate change, there is historic evidence of people behaving in an apparent PD-like pattern that is traditionally used to claim that PD models are explanatory. However, they argue, it is precisely this historical evidence that undermines the explanatory value of PD models, since it is the historical narrative itself that offers us more insights than the game-theoretic model by relating to a real causal chain (see Northcott and Alexandrova 2015: 76–77). If they are correct about PD models in general, similar worries may well apply to climate change mitigation.

Notably, an historic explanation for the problematic situation we find ourselves in now should include the fact that over long periods of time, when the industrialisation of economies took off, humanity was not aware of its environmental impacts. And for the periods when science began to grasp the dimensions of human impact on the climate, it might be more explanatory to analyse behaviour in terms of how inertia causes scientific insights to fail to be translated into political action, rather than taking willing exploitation of a public good to be the main (if not only) driving factor. Therefore, it follows that a lack of mitigation does not necessarily occur because of a special set-up of incentive structures assumed in the ToC, but can be caused by other historical, psychological or structural factors. Besides being potentially more explanatory, this leaves many options to avoid being trapped in some form of deterministic pessimism stemming from overconfidence in one specific model, and encourages us to look for alternative or complementary modelling options.

Thus, thinking of other historic causal explanations like these leads us to the next component of this strategy: coming up with alternative explanations, irrespective of whether non-trivial falsification attempts like the above can or will succeed.

3.2 Mapping out alternative explanation attempts

3.2.1 The Prisoner's Dilemma is not the only game in town

It has already been recognised that the ToC – and, as a more general form, the PD – is not the only available game-theoretic approach that aims to model the apparently failing climate negotiations we face. And just like the version of the ToC that was extended with prosocial preferences, not all of them yield the same daunting predictions while nevertheless employing potentially more

sensible assumptions. Consider the following example: rather than a single shot PD, it could be more accurate to portray climate change negotiations as an iterated PD, as people (or, for example, countries) make and change decisions about their emissions over time (see Wood 2011: 17–18). In extended models like this, many more Nash Equilibria are possible. In a Nash Equilibrium, the mutually best response to the action of other rational agents does not necessarily result in tragedy, but can sustain levels of cooperation. Also, allowing for behaviour like moral punishment in one's models allows for predictions of cooperation (see Boyd and Richerson 1992).

Hence, including plausible extensions of the ToC model design has major effects on its predictions. That does not necessarily make them better models, but at least it shows that the basic ToC model is not the only game in town. Essentially, this shifts the burden of proof to strict ToC proponents: if one were still to commit to the ToC, one would need to put forward excellent reasons for why the ToC model should be relied on; that is, to show that previously mentioned factors such as moral punishment, decision-making over time or prosocial preferences are indeed irrelevant. Literature suggests that their effect on climate change mitigation is considerable (see e.g. Ostrom 2009) and hence that they should be accounted for and not idealised away for the sake of simplicity. Here, complexity matters!

Although we believe that this is a promising approach, spelling out a complete alternative explanation is neither the aim of this paper, nor do we think that such a project will result in a singular model. As we argued above, we should not expect 'one perfect' model. We hold that models in principle are empirically underdetermined and that we may expect a whole palette of different models for various aspects of the problem to be necessary. Thus, we will consider the next candidate from our underdetermined palette, for which we aim to provide a non-exhaustive set of examples.

3.2.2 Decoupling wellbeing and GHG emissions

The major assumption that comes along with the ToC framing is that the individual payoff of the agents correlates with their GHG emissions. Because emitting is the dominant choice for every agent involved, individual wellbeing is assumed to be closely coupled with individual GHG emissions and choosing to emit yields a higher individual payoff, no matter what. And historically, this coupling seems to be well supported empirically, for example by the fact that Gross Domestic Product (GDP) is strongly correlated with GHG emissions (see Osobajo et al. 2020).

However, whether this connection will hold in the future or whether emissions and wellbeing can conceivably be decoupled remain open and hotly debated questions. If they can come apart, the payoff structure of the game we are in might not constitute a ToC, as the individual payoff would not

necessarily depend on individual emissions anymore. Plausible answers to this question also crucially depend on the employed proxy for societal wellbeing, and GDP does not in many cases appear adequate (see Ward et al. 2016).

What we can say with a high degree of confidence, however, is that several economic and technological developments offer help to transition to a GHG-decoupled wellbeing; even today, investors (households and businesses alike) have immediate monetary incentives to invest in low-GHG-emitting activities. As noted above, for example, renewable energy has in many cases become cheaper than fossil energy. Similar developments can be seen in other emissions-intensive industries such as mobility: the total cost of ownership of battery electric vehicles, for example, is lower than that of its combustion engine predecessors (see Hagman et al. 2016).

Whether GHG savings on products and services like these subsequently lead to a net reduction of emissions, however, is an even more controversial debate, as a decrease in cost (or increase in monetary incentives) may easily foster an overall increase in consumption and hence backfire. Nevertheless, focusing on the very existence of positive incentives to mitigate GHG emissions prompts further thoughts about us actually having a preference for agreement.

3.2.3 Rational preference for agreement

Framing climate change as a ToC implies that it is individually rational to emit GHG. On the other hand, it could also be argued that it constitutes a major rational incentive to create binding agreements in order to limit emissions. After all, collective ecological precaution is utility-maximising, as allowing climate change damage to happen constitutes a collective decrease in welfare. And while this general relation between individual and collective utility is admittedly captured in the ToC payoff structure, it may leave out many immediate benefits for the individual agent in question, whether a nation or an individual. These benefits, however, have the potential to move us away from the payoff structure of a ToC and hence make it individually rational to cut emissions.

It is precisely because of the notion of GHG mitigation being individually utility-maximising that we see people like Larry Fink, head of the world's largest asset manager BlackRock, forecasting that climate change is 'driving a profound reassessment of risk and asset values. And because capital markets pull future risk forward, we will see changes in capital allocation more quickly than we see changes to the climate itself' (Fink 2020: n.p.). This notion becomes even stronger once welfare and GHG emissions can be decoupled, as suggested in the previous section: not only will 'green' behaviour lower the risk of damage, but it may also simply be cheaper to use a non-emitting alternative.

Traditional explanations in the ToC framework for why the prospect of potentially gaining individual payoffs from mitigation (even in one's own lifetime) does not lead to a significant reduction of emissions often have to do with

the time-preferences of people. Accordingly, future payoffs are said to be heavily discounted by individuals (see e.g. Weitzman 2007), making short-sighted behaviour in PD-like patterns possible in the first place.

If we were to follow Fink's train of thought, however, his perspective constitutes a significant shift in what one might think to be the payoff structure of a climate change mitigation game; from this point of view, you do not have to be a climate activist who supports costly, 'irrational', economy-burdening policies to support climate change mitigation. We can simply argue from the viewpoint of an investor seeking to minimise risks for their investments, as 'climate risk is investment risk' (Fink 2020: n.p.). Furthermore, it appears to be this very line of argument that is commonly employed by environmentalists, economists and politicians when promoting environmental protection in public discourse, suggesting that this preference is truly present. Climate change mitigation, as such, can be considered perfectly consistent with individual utility maximisation and there seems to be little reason to restrict ourselves to the assumptions of the ToC. Outlining climate change in this manner has some important implications.

The first is that it leads to the assumption that we genuinely possess the preference for reaching an agreement, over failing to do so. If this is true, then the game we are in is different to the ToC, as this very preference has to be implemented in the payoff structure. Instead of a PD, climate change negotiations might then be better described by coordination games such as the Stag Hunt (SH) game. Both the PD and SH exhibit the classic free-rider problem, in which the worst outcome for a player is to reduce emissions while the other player happily continues to emit. The important difference is that in the SH the highest possible individual payoff is achieved through cooperation, while in the PD the highest possible individual payoff is achieved when free-riding. The question of which model better describes climate change is therefore a question of how severe the risks of climate change are considered to be, that is, how much we would value cooperation over free-riding. If one views climate change as an existential threat to humanity with a non-zero chance of civilisation collapse even when only a few nations choose to defect, then climate change is an SH, rather than a PD. If, in contrast, agents gain more utility from technological and economic domination instead of from lowering an existential threat for the whole group, then the payoff matrix of climate change is that of a PD (see DeCanio and Fremstad 2013: 182). This comparison between the PD and SH shows that it is not necessarily a particularly unlucky set-up of payoffs that leads to the historic and potential future mitigation failure. The main driving factor for our inaction could likewise be simply the '*failure of the leading governments to grasp the seriousness of the climate risk*' (Ibid.; emphasis in original). At this point, we can consider the choice between the PD and SH as largely underdetermined by the empirical evidence.

The second implication is that this approach renders the very lack of mitigation we observe today not as a result of rational behaviour, but instead as blatantly irrational. Climate change mitigation can be consistent with individual utility maximisation.

This prompts the question: is it justified to think that we as individuals are acting irrationally when it comes to climate change? Here, the behavioural sciences may offer us an insightful answer.

3.2.4 The behavioural sciences to the rescue

If we accept that non-mitigation exemplifies irrational rather than rational behaviour, a descriptive model for a causal explanation of current mitigation failure would have to aim to answer why and how this irrational behaviour came about. A plausible story would presumably leave behind the realms of pure game theory and incorporate insights from psychology and sociology, as well as political and historical science – or from the behavioural sciences in general.[6]

One such potential explanation, for example, is found in the prevalence of *cognitive biases*, as identified by Tversky and Kahneman (1973). Their laboratory experiments suggest that people often estimate probability or frequency with the help of simple judgemental heuristics. One of them, for instance, is the availability heuristic, by which people estimate the probability of an event by the ease with which instances of it come to mind. While such heuristics allow for fast decision-making, they also systematically violate basic rules of logic and probability theory, leading to biased and irrational judgements and behaviour.

A fitting example of how the availability bias may affect climate change mitigation behaviour might be the recent COVID-19 pandemic. In a matter of weeks after the virus first emerged, almost all nations closed their borders and public life came to a complete halt. Thus, apparently, if the danger and risk are sufficiently experienced, felt and perceived, drastic global political action and cooperation are possible.[7]

The potential aggregate damage brought upon us by climate change, however, is arguably significantly higher than the danger posed by one single pandemic. So what is the difference? Drawing on the availability bias might pave the road for a causal psychological explanation, insofar as the damages through climate change are a timely and geographically dispersed phenomenon, whereas COVID-19 is a more immediate threat. Hence, the danger of a pandemic is more available for subjective judgement than the danger of a

6. Note that this approach also corresponds with the argument by Northcott and Alexandrova (2015) mentioned earlier, according to which a Prisoner's Dilemma alone is not explanatory.

7. We are explicitly not arguing that the whole global response to COVID-19 was rational, we merely want to highlight that the global response was more drastic for the lesser threat.

seemingly more distant threat. Consequently, political action is more drastic for the less dangerous threat, which, if formulated like that, seems irrational.

Complementing the previous section, where we identified that climate inaction may conceivably be modelled as a form of irrational behaviour, contrary to the ToC story, behavioural sciences provide a method of explanation that pure game theory is lacking: causal explanations that transcend mere as-if conjectures. In social and environmental psychology, for example, there has been research along a similar line of thought long before COVID-19, known under the name 'value–action gap', which denotes a mismatch between valuing a stable climate on the one hand and inaction to sustain it on the other hand (see, for example, Kollmuss and Agyeman 2002).

4. HOW TO AVOID THE THEORY ROULETTE

Even though the ToC is often considered something of an obvious 'no-brainer' when it comes to climate change, we have seen that it is far from the only explanation one might possibly think of, and we have mapped out some of the conceivable options in the previous section. Further, we have also shown how this set of non-trivially unfalsified explanations can be considered underdetermined by providing a non-exhaustive set of explanation attempts. And, to the sceptical reader, we also want to point out that this proposed argument of underdetermination may hold, even when only a few or even none of our selected proposed explanation examples in the previous sections seem convincing.

Therefore, as long as it is not clear which of the available descriptive approaches is acceptable, we are left with an active choice about which explanatory frame to use when communicating the challenges ahead. This choice, as Quine suggested, is necessarily a pragmatic one (without thereby making it unscientific).

Additionally, and as Kopec pointed out himself, whatever explanation we choose can be expected to be performative if it fits the conditions spelled out in the introduction. This can create a positive feedback loop between the act of modelling and the gathering of evidence, after which the world is made to fit the model, rather than the other way around.

This conjunction of underdetermination and self-fulfilling performativity sets the stakes high for this particular choice among explanations. In this section, we aim to show that as long as we cannot rule out the most pessimistic self-fulfilling models (by non-trivial falsification), we should put the emphasis on the more optimistic ones, both in research and in communication. That is because whenever self-fulfilling performativity and underdetermination hold, we are in the midst of playing a collective form of Russian Roulette – just not with a cartridge, but with theories, of which one is pushing a catastrophe for humanity itself.

Spinning the cylinder of the revolver is appreciating and recognising underdetermination: this is because there are striking reasons to suspect that there is more than one single applicable model and, additionally, because it still remains uncertain which model is the right one. Because this choice is underdetermined it is a pragmatic choice and thus guided by relatively accidental reasons.

Pulling the trigger is publicising the model, spreading of the word and watching performativity happen. If we hit a performative deterministic-pessimist model, we are shooting ourselves dead. If it is neither pessimist nor performative, nothing bad will result from our model choice. The pivotal difference between the Russian Roulette analogy and our exposure to underdetermination and performativity is that we can actively choose not to load our revolver with a deadly cartridge – no sane person really wants to play Russian Roulette.

Instead of talking about a tragedy that is allegedly inevitable if everybody acted rationally or, even worse, about the tragedy being a rational necessity, which would only take in insights from one single underdetermined candidate model, a recognition of underdetermination gives us the opportunity to turn things around. As a start, this could be to explicitly frame mitigation and pushing for cooperation as being the utility-maximising thing to do. That includes pointing out forms of ignorance about the dangers and utility damages of climate change as an irrational cognitive bias.

Indeed, we may also want to make use of the performative nature of models, even if the assumed underlying subjective payoff structure would not dramatically change after such a switch in framing. We already have empirical evidence that, for instance, naming a situation differently without changing the payoff structure has effects on behaviour: in the infamous paper ‘The name of the game’, Liberman and colleagues conducted the same experiment twice, giving it two different names, Wall Street Game and Community Game. Even though it was the same Prisoner’s Dilemma on paper, the test subjects cooperated much more in the latter game (Liberman, Samuels and Ross 2004). What would happen, then, if the name of the game of climate change was not the ToC, but something that does not necessitate the largest collective action failure in human history? On grounds of performativity and the need for our model to help us better address the climate change crisis, we may very well want to call our collective action problem something like ‘The Human Extinction Challenge’, which makes the problem seem less like an insurmountable tragedy and more like a mutually beneficial task to prevent existential risk.

Finally, the fact that we can opt out of playing this form of collective Theory Roulette can be summarised in the form of a decision matrix: if underdetermination forces us to make a pragmatic choice, then emphasising non-ToC explanations is the dominant choice (Table 1).

State of the world Choice options	**ToC resembles an important mechanic and is applicable**	**ToC does not resemble an important mechanic and is not applicable**
Choose ToC in communication (load a cartridge)	*Inescapable apocalypse (as predicted)*	*Deal with performativity, which leads to potential apocalypse. Miss out on insights from other behavioural sciences*
Do not choose ToC in communication (do not load a cartridge)	*Delayed apocalypse (with performativity potentially in humanity's favour)*	*No apocalypse. More insights from more explanatory models*

Table 1. Nobody wants to play Theory Roulette; or: let's not load a cartridge. Source: author.

5. RESPONDING TO POTENTIAL OBJECTIONS

5.1 The ToC has a normative function

The first potential objection we want to address is the argument that models like the ToC often serve a normative function, not a purely descriptive one, as in 'we ought to collaborate'. Hence it is often used to show precisely why agreeing on coordinative action is rational. It is also this very notion that seems to be at play in the IPCC executive summary previously mentioned (IPCC 2014: pp. 211ff). There are two major problems with this objection, however.

The first has been pointed out by Northcott and Alexandrova (2015) with regard to the PD. That is, such normative advice can only be good if climate change is indeed accurately described as a PD – which is, as demonstrated, not only underdetermined but also highly challengeable. Otherwise, people would behave differently than predicted in the model anyway and the advice would miss its target. Since the ToC is merely a special version of the PD, the same conclusion holds for it.

The second problem is that to express a normative 'we ought to collaborate' is to express that one does value achieving cooperation over failing to do so. So, in any situation where this very preference is expressed, the SH approach mentioned earlier may serve as a better fit to the real-world payoff matrixes of people. Subsequent failure to act according to these preferences can then be explained not in the realms of game theory, but rather in terms of irrational cognitive biases like the aforementioned availability bias or the value–action gap.

5.2 This is wishful thinking

One might argue that such an argument concluding that we should emphasise an optimistic model is essentially wishful thinking, because a model is chosen simply on the grounds of whether one likes its implications or not, ignoring any kind of empirical model–world-fit. Obviously, this would be bad scientific practice. This objection, however, falls short in at least three ways.

First, since the ToC is underdetermined by empirical data, we are inevitably forced to rely on pragmatic criteria – which, as philosophers of science have elegantly argued, need not be seen as unscientific (see Reiss and Sprenger 2020). Being confronted with performativity in a situation of existential risk simply demands a precautionary approach, especially when forms of pragmatism in science are inescapable. A precautionary approach will use a less pessimistic model to bring the world closer to a more solvable situation.

Second, if the ToC is truly self-fulfilling, which is why we need this emphasis in the first place, then it already has obvious major flaws as a descriptive model, which we consider reason enough to justify looking at and emphasising both these flaws as well as alternative or more refined explanations.

Third, as we argued in section 3, we need to respond to underdetermination with model pluralism (see Veit 2020). That implies that without even waiting for *ex-post* (in)validation, we can *ex-ante* assume that a single model like the ToC will not suffice as a descriptive behavioural model. We can expect that multiple models for multiple aspects are needed simply by virtue of our epistemic uncertainty and the complexity of the real-world situation.

Despite having received a plethora of criticisms as the paradigm model of climate change mitigation failure (some of which we have mentioned: lack of evidence, historic explanations), the ToC might have received a disproportionate modelling commitment that seems rather unhealthy when looking at the more optimistic and potentially even more explanatory alternatives on which we could have focused. Thus, pushing to explore other explanations is likely to be epistemically beneficial in unforeseen ways, regardless of whether this push stems from a deliberate emphasis on non-ToC explanations based on normative grounds or from some other pragmatic criterium which we will necessarily employ anyway.

6. CONCLUSION

As this article aimed to show, we agree with many of Kopec's main conclusions regarding the danger of employing a self-fulfilling prophecy in climate change mitigation and the need for a different kind of framing. However, we strongly disagree with the arguments and reasons characterising his strategy for employing said alternative framing. Too often, weak arguments are accepted for

the sake of good conclusions. And weak arguments will subsequently make an adequate response much more difficult, as we have shown here.

If behavioural models regarding climate change mitigation suffer from performativity and deterministic pessimism, we agree with Kopec that we are essentially shooting ourselves in the foot. This is unfortunate because a proper response is urgently needed. However, we have shown that his strategy of emphasising the mere fact that our modelling assumptions are strictly speaking false is trivial. Thus it is unlikely to be convincing or helpful. Instead, what is urgently needed is a strategy that allows for the generally accepted margins of falsehood in scientific practice, i.e. considerations of idealisation, robustness and the evaluation of alternative modelling strategies.

Enter the underdetermination of scientific theory: for our case, it provides a rationale that justifies employing and looking out for alternative explanations, even when there is high confidence in the model. Just like with other complex phenomena, such as COVID-19 where multiple models are employed for different aspects, we urge epistemic humility and a more pluralistic approach that emphasises the limits of single models (see Veit, Brown and Earp 2021). Additionally, the existence of promising non-trivial falsification attempts gives further reason to consider abandoning the ToC as the paradigm model for climate change mitigation. Finally, we imagined the current conjunction of underdetermination and performativity as a Russian Roulette. This aims to provide pragmatic normative criteria for choosing between models when faced with underdetermination, as in our case: playing Theory Roulette is suicidal and choosing less pessimistic but likewise underdetermined options is the dominant choice.

As such, we see several advantages that the strategy proposed here has over Kopec's initial suggestions. First, it bypasses an impractical and trivial emphasis on strictly speaking false assumptions. Second, by putting the ToC in the broader context of underdetermination it shows that employing a particular model is an active choice. This is also the reason why this proposed strategy goes beyond the other two additional strategies that Kopec put forward, but which we have not yet mentioned explicitly: pointing out that other more optimistic explanation attempts exist, and encouraging those in the field to examine alternatives to the ToC (Kopec 2017: 13–15). Third, it provides both a useful catchphrase to communicate that we are forced to make that choice, as well as a rationale for making that decision.

Depending on how convincing one finds the attempts at non-trivial falsification and alternative explanations, this choice is quite an easy one. This is mainly because game theory has inevitable limits in what it can achieve, despite having proven itself to be a handy and important tool in many areas. This goes against the 'guiding prejudices of contemporary game theory', as game-theorist Herbert Gintis puts it, of game theory being 'sufficient to explain all of human social existence' (Gintis 2009: xiii). Even though the ToC provides

a neat story to portray a possible mechanism of free-riding and mutual exploitation, it does not say anything about the specifics. When applied to concrete real-world examples like climate change, the ToC appears to lack empirical grounding and arguably only offers a shallow explanatory power.

Furthermore, framing mitigation failure not as a necessity of rationality but as an irrational cognitive bias not only sheds light on other potential behavioural mechanics that might likewise be at play, but also helps to communicate the immediate benefits of climate change mitigation. Indeed, there has already been research on how a 'nudging' of that sort might proceed (see Andor and Fels 2018).

Although climate change has become the alleged obvious paradigm case of a ToC that is mentioned in executive summaries and introductory courses to economics alike, in the light of these criticisms and considering that we are in the midst of playing Theory Roulette, the ToC should arguably rather be the paradigm case of how oversimplistic game-theoretic models run the risk of being overrated and employed for invalid inferences about the real world. If it is true that, as Nicholas Stern puts it, climate change constitutes 'the greatest market failure the world has ever seen' (Stern 2007: viii) then it exemplifies the greatest challenge for the behavioural sciences to fathom why and how humanity stands in its own way to alleviate a dire existential catastrophe. Game-theoretic heuristics, as it stands, can only be a part of that puzzle.

After all, if we take performativity seriously, a commitment to the ToC model would be akin to voluntarily playing a Theory Roulette involving existential risk. It is on us to improve our chances by deliberately choosing other existing frameworks of explanation. Merely pointing to strictly speaking false assumptions will not suffice to make that choice. If underdetermination forces us to spin the Theory Roulette, we can at least avoid the risk of 'shooting ourselves'. These kinds of risk evaluations should not be seen as a scientifically spooky interference of non-epistemic values even to the most hardened empiricist. The avoidance of a human catastrophe can be an (all-)important virtue of a theoretical model.

ACKNOWLEDGEMENTS

We would like to thank Uwe Czaniera, Heather Browning, Damian Dibelius, Leander Schneider, Paulina Albert, Fiona Bauer and Timo Häcker in addition to two anonymous reviewers for their feedback on our manuscript. Furthermore, we would like to extend our gratitude to the audience at the 8th LSE-UBT Student Philosophy Conference, especially Carlos Nuñez Jimenes and Patricia Rich.

REFERENCES

Ackermann, K. and R. Murphy. 2019. 'Explaining cooperative behavior in public goods games: How preferences and beliefs affect contribution levels'. *Games* **10** (1): 15. **Crossref**

Andor, M. and K.M. Fels. 2018. 'Behavioral economics and energy conservation – a systematic review of non-price interventions and their causal effects'. *Ecological Economics* **148** (C): 178–210. https://econpapers.repec.org/article/eeeecolec/v_3a148_3ay_3a2018_3ai_3ac_3ap_3a178-210.htm.

Aydinonat, N.E. 2018. 'The diversity of models as a means to better explanations in economics'. *Journal of Economic Methodology* **25** (3): 237–251. **Crossref**

Boyd, R. and P.J. Richerson. 1992. 'Punishment allows the evolution of cooperation (or anything else) in sizable groups'. *Ethology and Sociobiology* **13** (3): 171–195. **Crossref**

Capros, P., A. De Vita, N. Tasios, P. Siskos and M. Kannavou. 2016. *EU Reference Scenario 2016. Energy, Transport and GHG Emissions Trends to 2050.* https://ec.europa.eu/energy/sites/ener/files/documents/20160713%20draft_publication_REF2016_v13.pdf (accessed 8 February 2022).

Cartwright, N. 2019. *Nature, the Artful Modeler: Lectures on Laws, Science, How Nature Arranges the World and How We Can Arrange It Better.* The Paul Carus Lectures 23. Chicago: Open Court.

DeCanio, S.J. and A. Fremstad. 2013. 'Game theory and climate diplomacy'. *Ecological Economics* **85**: 177–187. **Crossref**

Douglas, H. 2003. 'The moral responsibilities of scientists (tensions between autonomy and responsibility).' *American Philosophical Quarterly* **40** (1): 59–68. http://www.jstor.org/stable/20010097.

Edenhofer, O., N. Bauer and E. Kriegler. 2005. 'The impact of technological change on climate protection and welfare: Insights from the model MIND'. *Ecological Economics* **54** (2–3): 277–292. **Crossref**

Fink, L. 2020. 'The power of capitalism: Letter to CEOs'. https://www.blackrock.com/corporate/investor-relations/larry-fink-ceo-letter (accessed 19 January 2022).

Gintis, H. 2009. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, NJ: Princeton University Press.

Hagman, J., S. Ritzén, J. Janhager Stier and Y. Susilo. 2016. 'Total cost of ownership and its potential implications for battery electric vehicle diffusion'. *Research in Transportation Business & Management* **18**: 11–17. **Crossref**

Hardin, G. 1968. 'The Tragedy of the Commons'. *Science* **162** (3859): 1243–1248. **Crossref**

IPCC. 2014. 'Climate change 2014 – mitigation of climate change: Working Group III contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change'. https://www.ipcc.ch/site/assets/uploads/2018/02/ipcc_wg3_ar5_full.pdf (accessed 19 January 2022).

Kitcher, P. 2001. *Science, Truth, and Democracy.* Oxford, New York: Oxford University Press.

Kline, R., N. Seltzer, E. Lukinova and A. Bynum. 2018. 'Differentiated responsibilities and prosocial behaviour in climate change mitigation'. *Nature Human Behaviour* **2** (9): 653–661. **Crossref**

Kollmuss, A. and J. Agyeman. 2002. 'Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior?' *Environmental Education Research* **8** (3): 239–260. **Crossref**

Kopec, M. 2017. 'Game theory and the self-fulfilling climate tragedy'. *Environmental Values* **26** (2): 203–221.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Third edition. Chicago: The University of Chicago Press.

Levins, R. 1966. 'The Strategy of Model Building in Population Biology'. *American Scientist* **54** (4): 421–431. http://www.jstor.org/stable/27836590.

Liberman, V., S.M. Samuels and L. Ross. 2004. 'The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves'. *Personality & Social Psychology Bulletin* **30** (9): 1175–1185. **Crossref**

Longino, H.E. 1996. 'Cognitive and non-cognitive values in science: Rethinking the dichotomy'. In J. Nelson (ed.), *Feminism, Science, and the Philosophy of Science*, pp. 39–58. Synthese Library 256. Dordrecht: Springer Netherlands.

Mackenzie, D. 2006. 'Is economics performative? Option theory and the construction of derivatives markets'. *Journal of the History of Economic Thought* **28** (1): 29–55. **Crossref**

Mackenzie, D. 2008. *An Engine, Not a Camera: How Financial Models Shape Markets.* Cambridge, MA: MIT Press.

MacLean, D. 2015. 'Prisoner's dilemmas, intergenerational asymmetry, and climate change ethics'. In M. Peterson (ed.), *The Prisoner's Dilemma*, pp. 219–242. Cambridge: Cambridge University Press.

Mäki, U. 2009. 'Realistic realism about unrealistic models'. In H. Kincaid and D. Ross (eds), *The Oxford Handbook of Philosophy of Economics*, pp. 68–98. Oxford: Oxford University Press.

Naam, R. 2013. *The Infinite Resource: The Power of Ideas on a Finite Planet.* Lebanon, NH: University Press of New England.

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

Osobajo, O.A., A. Otitoju, M.A. Otitoju and A. Oke. 2020. 'The impact of energy consumption and economic growth on carbon dioxide emissions'. *Sustainability* **12** (19): 7965. **Crossref**

Ostrom, E. 2009. *A Polycentric Approach for Coping with Climate Change*. The World Bank. https://openknowledge.worldbank.org/bitstream/handle/10986/4287/WPS5095.pdf (accessed 19 January 2022).

Parker, W.S. 2020. 'Model evaluation: An adequacy-for-purpose view'. *Philosophy of Science* **87** (3): 457–477. **Crossref**

Patt, A. 2017. 'Beyond the tragedy of the commons: Reframing effective climate change governance'. *Energy Research & Social Science* **34**: 1–3. **Crossref**

Poteete, A.R., E. Ostrom and M. Janssen. 2010. *Working Together: Collective Action, the Commons, and Multiple Methods in Practice.* Princeton, NJ: Princeton University Press.

Quine, W.V.O. 1951. 'Main trends in recent philosophy: Two dogmas of empiricism'. *The Philosophical Review* **60** (1): 20–43. **Crossref**

Reiss, J. and J. Sprenger. 2020. *Scientific Objectivity*. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 edition). https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/ (accessed 19 January 2022).

Stanford, K. 2017. 'Underdetermination of Scientific Theory'. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition). https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/ (accessed 19 January 2022).

Stern, N. 2007. *The Economics of Climate Change: The Stern Review.* Cambridge: Cambridge University Press.

Tilman, A.R., A.K. Dixit and S.A. Levin. 2019. 'Localized prosocial preferences, public goods, and common-pool resources'. *Proceedings of the National Academy of Sciences of the United States of America* **116** (12): 5305–5310. **Crossref**

Tversky, A. and D. Kahneman. 1973. 'Availability: A heuristic for judging frequency and probability'. *Cognitive Psychology* **5** (2): 207–232. **Crossref**

Veit, W. 2020. 'Model pluralism'. *Philosophy of the Social Sciences* **50** (2): 91–114. **Crossref**

Veit, W. 2021. 'Model diversity and the embarrassment of riches'. *Journal of Economic Methodology* **28** (3): 1–13. **Crossref**

Veit, W., R. Brown and B.D. Earp. 2021. 'In science we trust? Being honest about the limits of medical research during COVID-19'. *The American Journal of Bioethics* **21** (1): 22–24. **Crossref**

Ward, J.D., P.C. Sutton, A.D. Werner, R. Costanza, S.H. Mohr and C.T. Simmons. 2016. 'Is decoupling GDP growth from environmental impact possible?' *PLoS ONE* **11** (10): e0164733. **Crossref**

Weisberg, M. 2015. *Simulation and Similarity: Using Models to Understand the World.* Oxford: Oxford University Press.

Weitzman, M.L. 2007. 'A review of the Stern Review on the economics of climate change'. *Journal of Economic Literature* **45** (3): 703–724. **Crossref**

Wood, P.J. 2011. 'Climate change and game theory'. *Annals of the New York Academy of Sciences* **1219**: 153–170. **Crossref**