Please cite as: Ortmann, J. & Veit, W. (2021). Theory Roulette: Choosing that Climate Change is not a Tragedy of the Commons. *Preprint*.

Check https://walterveit.com/publications/ for citation details once published.

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

Jakob Ortmann and Walter Veit

Abstract

Climate change (CC) has become a paradigm case for externalities in general and for the *Tragedy of the Commons* (ToC) model by Hardin in particular. This is worrying as we have reasons to suspect that models like ToC are performative, such that they might become self-fulfilling prophecies. In this paper, we aim to enhance a strategy proposed by Matthew Kopec to cope with the self-fulfilling nature of ToC. First, we show how Kopec's strategy about emphasising that ToC relies on strictly speaking false assumptions is unlikely to be a successful strategy. To construct a more promising strategy we argue that the argument of underdetermination implies that the employment of a specific model is an active choice that is guided by pragmatic criteria. Furthermore, picturing underdetermination in the case of CC as a form of Russian Roulette provides a rationale to choose between these underdetermined models.

Keywords: Tragedy of the Commons, Climate Change, Philosophy of Science, Game Theory, Ethics of Climate Change

CONTENT

1 FRAMING THE PROBLEM

Climate Change is often modelled as a *Tragedy of the Commons* (ToC). Indeed, this has happened so many times that it seems to have evolved into a paradigm example for game theory and microeconomics – the ultimate tragedy of the commons: the prisoner's dilemma (PD) of doom.[1] Moreover, there is a high degree of confidence in modeling climate change this way which, for instance, lead the Intergovernmental Panel on Climate Change (IPCC) to ascribe "high confidence" (IPCC 2014, p. 211) to the correctness of that model.

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 challenge. This concern mainly arises because historic methods to dissolve the ToC appear to not apply to greenhouse gas (GHG) emissions, (see Kopec, M. 2016, p. 6) which gives rise to a class of worries that we will condense as *deterministic pessimism*. Matthew Kopec, referring to this concern, was inclined to frame our current situation as "[…] the climate crisis that seems rationally forced upon us", (Kopec, M. 2016, p. 15) and Hardin, the originator of the term 'Tragedy of the Commons', accordingly framed the tragedy in terms of a lack of a "technical solution" (Hardin, G. 1968, p. 1248): meaning, that if we are sufficiently accurately portrayed by rational, utility-maximising agents assumed in the ToC model, we are faced with an unsolvable problem. This alone provides reason enough to review what assumptions are explicitly and implicitly stipulated in detail when we employ ToC to descriptively model our failure for sufficient mitigation.

Worse perhaps, there is empirical evidence that some behavioural models are performative and therefore apt to become self-fulfilling prophecies. When used and implemented as a model in scientific or political discourse these models have a propensity to amplify and causally interfere with what they merely want to describe (see e.g. Mackenzie, D. 2006). The characteristics of 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 Kopec, M. 2016, p. 9f.; And see Mackenzie, D. 2006, p. 43ff.). According to Kopec, the ToC model applied to climate change is

[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, D. (2015). In this paper, however, we will treat ToC as a variation of PD with incremental decisions by more than one agents. Related terms are *commons problem*, *common pool resource problem* and *externalities problem.*

highly likely to satisfy these characteristics and thus is likely to be a self-fulfilling prophecy.

If this is true, asking the question of how confident we are about ToC is not only a purely positive-descriptive endeavour but one with severe and dangerous consequences of ethical dimension. Why, we ought to ask then, should we be confident that the ToC is the best way to model climate change negotiations? And if ToC turns out to be false, not applicable or just too vague to model our failure of mitigation, we will have even more urgent reasons to discard ToC as a misconception or oversimplified heuristic, before the misconception itself may aid to bring about the tragedy.

In an attempt to maintain at least some optimism about humanity's ability to alleviate the climate crisis as well as to avert the self-fulfilling performativity of the commons model, Kopec suggests some strategies we could employ when we communicate the climate crisis via that model. One of them is "[…] 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 how the assumptions of the model are not likely to be strictly speaking true" (Kopec, M. 2016, p. 12).

While we agree with Kopec's general point about the danger of 'performativity' or 'self-fulfillingness' in the case of ToC, we suggest that the solution strategy he proposes needs to be substantially improved. As we aim to show in section 2 by drawing on the philosophy of science literature, models in microeconomics tend to always be highly idealised – at best, they only represent a part of the world, perhaps a particular dynamic or process, and shed light on some conditional causal connections, i.e. if they succeed. As a result, talk of the *strictly speaking falsity* of ToC is trivial and thus unlikely to be helpful.

However, by recognizing that models are always underdetermined by empirical evidence, we are provided with an alternative solution, that is to rely on a more pluralistic approach to modeling complex phenomena, as it is common in other complex sciences (like, e.g., ecology or weather forecasts). But because alternative and underdetermined models provide vastly different predictions and explanations, a pragmatic framework on how to choose between these models is needed, and, perhaps more importantly, some guidance on how to effectively communicate them to the public. One suggestion of a requirement of such a pragmatic framework on communication emerges from a particular reading of underdetermination, that we will spell out in a subsequent section: The combination of both self-fulfilling performativity together with underdetermination

of behavioural theories renders us to be part of a kind of Russian Roulette that we shall dub *Theory Roulette*.

2 ON STRICTLY SPEAKING FALSE ASSUMPTIONS

On first sight, Kopec's strategy appears reasonable. Since the ToC is a model, it necessarily *must* employ idealised assumptions specifically to enable it to say anything useful (see e.g. Weisberg, M. 2015). And clearly, we ought to point out the boundaries and limitations of our models when we employ them, and even more so when we are worried about potentially dangerous performativity of our model. Kopec's strategy is also in line with a tradition of criticising economic models like ToC for being too simple, unrealistic and ignoring important features of the real world. In light of contemporary philosophy of science, however, Kopec's strategy turns out to be rather hollow – for at least three reasons:

Firstly, it is unhelpful to add the clarification that the ToC is *strictly speaking* false, for it suggests that there is something especially problematic 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. Take, for example, the argument by the *notorious pessimistic meta-induction*, first formulated by Larry Laudan, L. (1981). If we take a retrospective on the history of scientific progress we are presented with an ongoing abandonment of formerly accepted theories and therefore should take seriously the idea that our current best theories will eventually suffer the same fate (see Laudan, L. 1981). Thus, the criterium of *strictly speaking* falsity can be considered to hold for almost any theory that we have. And still, yet, we rely on many of these *strictly speaking* false theories and models in many different ways. They are here to stay - and yet, despite being strictly speaking false, they are the best tools we have to understand reality. Because Kopec's emphasis on falsity – as formulated above – would thereby apply to each and every theory of any kind, this strategy turns out to be a rather trivial non-starter.

Also, consider the *argument of underdetermination*. According to the holistic account of that argument, the empirical data available at any time is insufficient to coercively decide between co-existing theories that are (1) compatible with a given set of observations and (2) mutually contradictory (see Stanford, K. 2017). Hence, the falsification of any theory leaves us with the open question of which remaining theory is the most reasonable. Willard Quine, originator and proponent of that account, thus concludes that "[…] the considerations which guide [someone] in warping his scientific heritage to fit his continuing sensory promptings are […] *pragmatic*." (Quine, W. 1951,

p. 43; emphasis added). The argument of underdetermination contributes to the problem at hand in two ways. First, it provides an additional reason to consider (strictly speaking) falsity to be common among our theories, as we cannot coercively choose between multiple mutually exclusive models. 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 come back to this second point at the end of this paper.

Finally, idealisation, and hence falsehood, is often considered an *appreciated feature, not a bug* of a model. Imagine, for example, standing in front of a subway map. 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 close as possible. The common abstractions of subway maps render them less approximative to the real world than more detailed maps. Nevertheless, they provide us with relevant and useful insights for your specific task by not obscuring these insights with information that is inessential to you (Weisberg 2015). Models or theories that include *all* the variables would simply be unusable. Variants of arguments along this line of false models still being explanatory have been defended in the past, for example, by Uskali Mäki with his account of models as isolations (Mäki, U. 2009; For an overview of other arguments see Weisberg, M. 2015).

For these three reasons (meta-induction, underdetermination and welcomed idealisation), only pointing out that ToC relies on *strictly speaking* false assumptions is unlikely to become a satisfactory strategy to convince people that ToC is, in fact, unhelpful and not shedding light on some relevant mechanism that leads to free riding. A proper strategy for alleviating a self-fulfilling climate change tragedy would have to be able to account for these three allowances of falsehood – and cannot rely on some imprecise emphasis on *strictly speaking* false assumptions.

Instead of drawing on the notion of strict truth of falsity (which is largely abandoned among contemporary philosophers of science), we propose a strategy that seeks to fracture the seemingly ubiquitous and potentially dangerous confidence of climate change as a ToC in a more convincing manner. This strategy is composed of two parts.

First, we will take the general problem of underdetermination literally for the case of climate change mitigation failure and roughly map out several potential attempts that likewise try to explain the situation we find ourselves in. Secondly, this underdetermination will prompt us with the need for pragmatic criteria to choose between these co-existing attempts of explanation. Here, we propose pragmatic framework to help us do exactly that, which we dub *Theory Roulette.*

Together, both parts form a strategy to alleviate a self-fulfilling tragedy that aims to go beyond a trivial emphasis on *strictly speaking* false assumptions.

3 UNDER(DETER)MINING TOC

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

For example, a minimal extension to the standard representation of ToC is to add some form of prosocial preferences to the assumed agents. One might be justified to believe 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. When prosocial preferences are present, the difficulty of solving collective action problems, like climate change, is generally thought to be substantially reduced (See e.g.Ackermann, K./Murphy, R. 2019; Kline, R. et al. 2018; Tilman, A.R./Dixit, A.K./Levin, S.A. 2019).

We are now prompted with two variants of ToC. Which one is climate change – ToC with or without prosocial preferences? This is underdetermination biting.

An answer to this particular question, 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 in both what their respective predictions really mean as well as what precise question they even attempt to resolve. 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, nation leaders, or all of them at the same time (e.g. IPCC 2014, p. 211). A more mature behavioural theory, however, would presumably have to employ considerable contrasting between each of these alterations.

Also, 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 pay off to attack ToC directly. That is to figure out where exactly the boundaries of the explanatory power of ToC lie and in which ways empirical evidence has failed to support the very implications that motivate deterministic

pessimism and are self-fulfilling. We will refer to this as attempting a *non-trivial* falsification (as opposed to relying on *strictly speaking* false assumptions). This attempt may turn out in two different ways:

First, if an attempt of non-trivial falsification succeeds, Kopec's strategy would have to be bolstered up in that ToC for climate change is not only *strictly speaking* false, but also plainly speaking. ToC would have to be abandoned and one could hope for other models with more optimistic predictions. If they indeed are optimistic, then self-fulfilling performativity could even be a welcome feature. In section 3.1, we will roughly sketch out various promising attempts that aim to achieve exactly that.

Second, if falsification does not succeed, then ToC is still among the underdetermined candidates of a descriptive model for climate change mitigation failure. We will map out a selection of these alternative candidates in section 3.2. In this second case, recognising that we are in the midst of playing a Theory Roulette may provide pragmatic criteria on how to proceed when faced with said underdetermination.

3.1 Attempting Non-Trivial Falsification

Since we have already seen that *strictly speaking* falsity is not an adequate property of a model to be considered false such that it is not employed for an explanation, an attempt of 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 presumably have to be an answer of degree which, in turn, highly depends on aspects like what phenomenon exactly under which conditions is the subject of a particular ToC portrayal.

As a starter, this would include disambiguating the earlier mentioned vagueness about who, precisely, the agents are supposed to be. Are our agents the representatives of nations, that failed to reach and enforce adequate agreements in Kyoto, Copenhagen and Paris? Or are they private households that seek to minimise their expenditure on power consumption? Or are they parents who rather use their air-conditioned car to drive their children to primary school because it seems more convenient and safe than to use a bicycle instead? In all these cases it has to be asked: Is ToC the best explanation we have?

A striking case to disentangle this complexity of potential ToC instantiations has been made by Eleanor Ostrom, already over a decade ago (see Ostrom, E. 2009). It is, what she calls, a polycentric approach. Her critical review of 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, E. 2009, p. 9). This directly corresponds to the earlier identified

vagueness about whom the agents are supposed to resemble, i.e. which scale we are looking at. Hence, according to Ostrom, it is not a good scientific approach to only look at one particular scale for costs and benefits of GHG mitigation, but instead at the multiplicity of effects of diverse actions on multiple scales and their reciprocating influence (Ostrom, E. 2009, p. 32ff.). It could be argued then, that by ignoring such multi-scale complexity, a critical degree of falsehood by idealisation is exceeded and ToC is indeed too simple to be explanatory – the same way a subway map that does not show you all the lines available exceeds a critical degree of idealisation for your specific purpose.

The second ground of Ostrom's criticism is the blatant lack of empirical evidence for the conventional ToC predictions. The unambiguous (and in the case of climate change, frightening) predictions are simply not supported by observation (Ostrom, E. 2009, p. 10). This insight cannot be overstated for a model of which the supposed paradigm case is the largest potential humanitarian crisis in history. Besides providing a book-length analysis of these empirical findings, Poteete, A.R./Ostrom, E./Janssen, M. (2010) call for an updated theory of collective action that accounts for diverse organising of commons at multiple levels. The upshot of this callout is "[…] that it 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, E. 2009, p. 39).

Another approach is to question the assumption of whether climate change mitigation does even meet the criteria for being a common pool resource. Various concerns about this crucial assumption (that is often taken for granted) have been raised e.g. by Anthony Patt, A. (2017). One of them is that there indeed do exist potential technical solutions that yield medium-term costs of eliminating GHG emissions to be trivial, if not negative (see Patt, A. 2017, p. 2). According to Patt, this insight 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, A. 2017, p. 2).

The third and last approach of non-trivially falsifying ToC we want to highlight is from Robert Northcott and Anna Alexandrova: straight-up refusing that a formal model like the Prisoner's Dilemma (PD) can be causally explanatory. According to them, historic evidence of people behaving in PD-like patterns is traditionally used to claim that PD models are explanatory. However, they argue that the very same historic examples become evidence *against* the explanatory and heuristic value of PDs: because

in these cases a historic explanation itself is much more insightful than a PD heuristic on top of that could ever be (see Northcott, R./Alexandrova, A. 2015).

The same applies, one might argue, to climate change. A historic explanation for the situation we find ourselves in could include, for example, 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 of when science began to grasp the dimensions of human impact on the climate, it might be more explanatory to analyse behaviour in terms of inertia of scientific insights to be translated in political action. Besides being potentially more explanatory, this leaves open many doors for not being trapped in some form of deterministic pessimism stemming from overconfidence in one specific model.

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

3.2 Mapping-Out Alternative Explanation Attempts

3.2.1 Prisoner's Dilemma Is Not The Only Game In Town

By far, ToC – and as a more general form: the Prisoner's Dilemma (PD) – is not the only available game-theoretic approach that aims to model the apparent climate negotiations we are faced with (see e.g.Wood, P.J. 2011). Moreover, just like ToC extended with prosocial preferences, not all of them yield the same daunting predictions, while nevertheless employing potentially more sensible assumptions. Selected examples would be the following:

Rather than a single shot PD, it could be more accurate to portray climate change negotiations as an iterated PD, as people (or countries etc.) make and change decisions about their emissions over time (see Wood, P.J. 2011, p. 17f.). In models like this, many more nash equilibria are possible, besides the tragedy. Also, allowing for behaviour like moral punishment in one's models allows for predictions of cooperation (see Boyd, R./Richerson, P.J. 1992).

Hence, including decision-making over time as well as moral punishment are alternations in ToC model design that have major effects on its predictions. That does not necessarily make them better fitting theories (e.g. they still work with a similar degree of idealisation) but at least they show that ToC is not the only game in town. If one were still to commit to ToC in an undogmatic manner one would need to put forward damning reasons of why ToC is still the obvious choice, i.e. to show that moral punishment, decision-making over time or prosocial preferences are indeed irrelevant factors.

3.2.2 Decoupling Wellbeing and GHG Emissions

Although arguments of this kind have often been subject to much scrutiny it still is unclear whether future societal wellbeing can be decoupled from GHG emissions. Possible answers to that question also heavily depend on the employed proxy for societal wellbeing (e.g. gross domestic product does in many cases not appear adequate) (see Ward, J.D. et al. 2016).

What we can say with certainty, however, is that at least several economical and technological advancements offer help to transition to decoupled wellbeing: already today investors (households and businesses alike) have immediate monetary incentives to invest in low-GHG-emitting activities.

For example, 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, P. et al. 2016). Also, for example, the total cost of ownership of battery electric vehicles is lower than that of its combustion-engined predecessors (see Hagman, J. 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 overall consumption and hence backfire. However, 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 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 there is, in fact, a stark rational incentive to reach an agreement to limit emissions. After all, collective ecological precaution is utility maximising, as allowing for damage through climate change constitutes a collective decrease in welfare. Traditional explanations in the ToC framework for why this does not lead to a significant reduction of emissions often have to do with time preferences of people, according to which future payoffs are heavily discounted (e.g. Weitzman, M.L. 2007), which is what makes short-sighted behaviour in PD-like patterns possible in the first place.

Because of the very notion of GHG mitigation being utility maximising, however, 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, L. 2020).

This 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 that supports costly, "irrational", economy-burdening policies to support climate change mitigation. Arguing from the viewpoint of an investor that seeks to minimise risks for her investments suffices as "climate risk is investment risk" (Fink, L. 2020). Furthermore, it appears to be this very line of arguing that is mostly employed by environmentalists, economists and politicians alike when promoting the case of environmental protection in public discourse, furthermore suggesting that this preference is truly present. Climate change mitigation, as such, can be considered perfectly consistent with utility maximisation – just this time a new important variable is added to the utility function assumed: the stability of climate.

Outlining climate change like this has some important implications: First, it motivates the consideration that we genuinely possess the preference to reach an agreement over failing to do so. If this is true, then the game we are in has changed. Instead of a prisoner's dilemma, climate change negotiations might be better described by a game-theoretic model like Battle of the Sexes (BoS) (see MacLean, D. 2015, p. 226).

Secondly, recognising that we are in a different kind of game such as BoS as well as acknowledging that climate action is consistent with individual utility maximisation, renders lacking mitigation not as a result of rational behaviour (which is an expression quite positively connotated which does not help to solve dangerous performativity) but as blatantly irrational instead. Irrationality, here, would denote the inability to live up to one's real preferences – in this case: to reach an agreement. Thus, in the next section, consider a sketched-out alternative explanation of climate change mitigation failure that explicitly frames the current mitigation failure as irrational: the availability bias.

3.2.4 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. An answer to this question would presumably leave behind the realms of pure game theory and incorporate

insights from psychology, sociology as well as political and historical science – or from the behavioural sciences, in general.[2]

One such potential explanation, for example, we find in the *availability bias*, coined by Tversky, A./Kahneman, D. (1973). A fitting example of this might be the recent COVID-19 pandemic. In a matter of weeks after the virus breakout, almost all nations closed 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 is possible.

The potential aggregate damage brought upon us by climate change, however, is arguably significantly higher than the danger posed by one single pandemic. Here we clearly underreact – so what is the difference? Drawing on the availability bias might insofar pave the road for a causal psychological explanation as the damages through climate change are a delayed and global phenomenon whereas COVID-19 is a more immediate threat. Hence, the pandemic is more available for subjective judgement of danger. Consequently, political action is more drastic for the less dangerous threat, which, if formulated like that, seems irrational.

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 e.g. Kollmuss, A./Agyeman, J. 2002).

4 THEORY ROULETTE

Even though ToC is often considered an obvious no-brainer when it comes to climate change, we have seen that it is by far not the only explanation one might think of – some of which we have mapped out in the previous section. We also hinted at how this set of non-trivially unfalsified explanations can be considered underdetermined. Therefore, as long as it is not clear which of these is the most reasonable descriptive approach, 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.

Additionally, and as Kopec pointed out, the explanation we choose can be expected to be performative. This conjunction of underdetermination and self-fulfilling performativity sets the stakes high for this particular choice among explanations. We aim to show that as long as we cannot coercively rule out the most pessimistic model, we are good advised to lay emphasis on the more optimistic ones, both in research and

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

in communication. That is because if self-fulfilling performativity and underdetermination hold, we are currently in the midst of playing a form of Russian Roulette – just not with a cartridge, but with theories, one of which being deadly. Spinning the cylinder of the revolver is appreciating and recognising underdetermination. Pulling the trigger is the spreading of the word and watching performativity happen. The pivotal difference of this analogy to our exposure to underdetermination and performativity is that we can actively choose not to load our revolver with a deadly cartridge.

Instead of talking about a tragedy that allegedly is 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, a recognition of underdetermination gives us the opportunity to flip the table. As a starter, this could be explicitly framing mitigation as being the rational and utility maximising thing to do. That includes pointing out forms of ignorance about the dangers and utility damages of climate change, of which we have an insurmountable amount of scientific evidence, 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 of 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 et al. (2004) conduct the same experiment twice just 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 one (Liberman, V./Samuels, S.M./Ross, L. 2004). What would happen, then, if the name of the game of climate change was not ToC, but something that does not necessitate the largest collective action failure in human history?

And lastly, this notion of us being prompted with an active choice to (not) let Theory Roulette perpetrate can be further condensed in form of a decision matrix: If underdetermination forces us to make a pragmatic choice, then emphasising non-ToC explanations is the dominant choice.

	ToC does resemble an important mechanic and is applicable	**ToC does not resemble an important mechanic and is not applicable**
Choose ToC in communication	*Inescapable apocalypse*	*Deal with performativity, which leads to potential apocalypse. Miss on insights of other behavioural sciences*
Choose not ToC in communication	*Delayed apocalypse with performativity in our favour*	*More insights through more explanatory behavioural sciences*

Figure 1. [Theory Roulette as a decision matrix].

5 RESPONDING TO POTENTIAL OBJECTIONS

5.1 Normative Function

The first potential objection we want to address is to say that models like ToC often serve a normative function, not a purely descriptive one, as in "we ought to collaborate". Hence, it is often used precisely to show why agreeing on coordinative action is rational. It is also this very notion that seems to be at play in the earlier mentioned IPCC executive summary (see IPCC 2014, p. 211ff). We see two flaws in this remark.

The first has been pointed out by Northcott and Alexandrova in regards to the Prisoner's Dilemma (PD). That is, such normative advice can only be good if climate change is indeed accurately described by a PD. Otherwise, people would behave differently than predicted in the model anyway and the advice would miss its target. "Thus a normative perspective offers no escape from the central problem, namely the ubiquitous significance in practice of richer contextual factors unmodeled by the Prisoner's Dilemma" (Northcott, R./Alexandrova, A. 2015, p. 84). Considering that ToC is merely a special version of the PD, the same holds for ToC.

Secondly, to express that 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 Battle of the Sexes (BoS) approach mentioned earlier does capture these preferences. Subsequent failure to act according to these preferences can then be explained not in the realms of BoS, but rather in terms of irrational cognitive biases like the availability bias or the value-action-gap.

5.2 This Is Wishful Thinking

One might argue that such a callout on emphasis, which is a normative claim about which explanation to propagate and which not, is essentially a form of wishful thinking and

hence bad scientific practice – because one deliberately chooses another model, simply depending on whether one does like its implications and not on any epistemic grounds such as non-trivial falsification.

This objection, however, falls short in multiple ways. First, because if ToC is underdetermined, we are prompted with a pragmatic choice of some sort anyway – which is a pure epistemic and not a normative "wishful" argument, albeit more subtle than a direct non-trivial falsification of ToC would be. Being confronted with performativity, to be precautionary and hence potentially prevent damage is a reasonable criterium to be incorporated in this pragmatic decision.

Secondly, if ToC is truly self-fulfilling, which is why we need this emphasis in the first place, then ToC 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.

Thirdly, contemporary philosophy of science seems to settle on the idea of at least some form of model pluralism (see Veit, W. 2020; 2021). That is to say that each phenomenon has multiple different aspects that require multiple different models for explanations. So, without even waiting for ex-post (in)validation, we can ex-ante assume that a single model like ToC will not suffice as a descriptive behavioural model after all. We can expect that at least multiple models for multiple aspects are needed simply in virtue of our epistemic uncertainty. Underdetermination requires pluralism both in method and models.

Considering its status as the one and only paradigm model of climate change mitigation failure, despite there being lots of substantial criticism (e.g. mentioned earlier: lack of evidence, historic explanations), we have here criticized the disproportionate focus of scientific effort and communication in that one direction – which seems rather unhealthy when looking at the more optimistic and potentially even more explanatory alternatives that we could have spent more time on. Thus, being pushed to explore other directions of explanations is likely to be beneficiary anyway in this case – regardless of whether this push stems from a deliberate emphasis based on normative grounds or from some other pragmatic criterium which we will necessarily employ anyway.

6 CONCLUSION

Starting with worries of deterministic pessimism and performativity regarding climate change being a ToC, we highlighted a strategy by Kopec that aims to deal with these concerns. We have also shown how a strategy relying on *strictly speaking* falsity in

assumptions is trivial and thus unlikely to be convincing. What is needed, is a strategy that allows for the generally accepted margins of falsehoods in scientific practice.

Enter underdetermination of scientific theory: for our case, it provides a rationale that justifies employing and looking out for alternative explanations. Additionally, the existence of strong non-trivial falsification attempts gives further reason to abandon ToC as a descriptive model for climate change mitigation.

Lastly, we pictured the conjunction of underdetermination and performativity as a Russian Roulette which aims to provide pragmatic normative criteria to choose between models when faced with underdetermination as in our case.

As such, we see several advantages to Kopec's suggestions in the strategy proposed here. First, it bypasses an impractical and trivial emphasis on *strictly speaking* false assumptions.

Secondly, by putting 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 additionally put forward but which we did not mention explicitly yet: pointing out that other more optimistic explanation attempts exist (see Kopec, M. 2016, p. 13ff.).

Thirdly, it provides both a catchphrase to communicate that we are necessitated to make that choice, as well as a rationale for making that decision. Depending on how convincing one does find the attempts of non-trivial falsification, this choice is quite an easy one, even on purely epistemic grounds. Because although it has proven to be a handy and important tool in many areas, game theory has boundaries. This is 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, H. 2009, p. xiii). Even though ToC provides a neat story to portray a possible mechanism of freeriding and mutual exploitation, it does not say a word about the specifics. When applied to concrete real-world examples, like climate change, the explanatory depth of ToC appears shallow, as argued by Northcott, R./Alexandrova, A. (2015).

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

Although climate change has become the alleged obvious paradigm case of 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, ToC should arguably rather be the paradigm case of how game-theoretic models are overrated and employed for invalid inferences about the real world. If it is true that, as Nicholas Stern puts it, climate change (emissions being a public good and climate change thus a ToC) constitutes "[…] the greatest market failure the world has ever seen" (Stern, N. 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 that should not be overstated.

After all, if we take *performativity* serious we better should win this match of Theory Roulette – it is on us to improve our chances by choosing other existing explanations and pointing on *strictly speaking* false assumptions will not suffice to make that choice.

7 REFERENCES

Ackermann, K./Murphy, R. (2019): Explaining Cooperative Behavior in Public Goods Games: How Preferences and Beliefs Affect Contribution Levels, in: Games, Vol. 10, No. 1, p. 15.

Andor, M./Fels, K.M. (2018): Behavioral Economics and Energy Conservation – A Systematic Review of Non-price Interventions and Their Causal Effects, in: Ecological Economics, Vol. 148, C, p. 178–210.

Boyd, R./Richerson, P.J. (1992): Punishment allows the evolution of cooperation (or anything else) in sizable groups, in: Ethology and Sociobiology, Vol. 13, No. 3, p. 171–195.

Capros, P./De Vita, A./Tasios, N./Siskos, P./Kannavou, M. (2016): EU Reference Scenario 2016. Energy, transport and GHG emissions Trends to 2050, in: https://ec.europa.eu/energy/sites/ener/files/documents/20160713%20draft_publication_REF2016_v13.pdf (accessed 25/01/2020).

Fink, L. (2020): Letter to CEO's, in: https://www.blackrock.com/corporate/investor-relations/larry-fink-ceo-letter (accessed 25/01/2020).

Gintis, H. (2009): The bounds of reason. Game theory and the unification of the behavioral sciences.

Hagman, J./Ritzén, S./Stier, J.J./Susilo, Y. (2016): Total cost of ownership and its potential implications for battery electric vehicle diffusion, in: Research in Transportation Business & Management, Vol. 18, p. 11–17.

Hardin, G. (1968): The Tragedy of the Commons, in: Science (New York, N.Y.), Vol. 162, No. 3859, p. 1243–1248.

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.

Kincaid, H./Ross, D. [Ed.] (2009): The Oxford handbook of philosophy of economics, Oxford.

Kline, R./Seltzer, N./Lukinova, E./Bynum, A. (2018): Differentiated responsibilities and prosocial behaviour in climate change mitigation, in: Nature human behaviour, Vol. 2, No. 9, p. 653–661.

Kollmuss, A./Agyeman, J. (2002): Mind the Gap: Why do people act environmentally and what are the barriers to pro-environmental behavior?, in: Environmental Education Research, Vol. 8, No. 3, p. 239–260.

Kopec, M. (2016): Game Theory and the self-fulfilling climate tragedy, in: Environmental Values.

Laudan, L. (1981): A Confutation of Convergent Realism, in: Philosophy of Science, Vol. 48, No. 1, p. 19–49.

Liberman, V./Samuels, S.M./Ross, L. (2004): The name of the game: predictive power of reputations versus situational labels in determining prisoner's dilemma game moves, in: Personality & social psychology bulletin, Vol. 30, No. 9, p. 1175–1185.

Mackenzie, D. (2006): Is Economics Performative? Option Theory and the Construction of Derivatives Markets, in: Journal of the History of Economic Thought, Vol. 28, No. 1, p. 29–55.

MacLean, D. (2015): Prisoner's Dilemmas, intergenerational asymmetry, and climate change ethics, in: Peterson, M. [Ed.] (2015): The prisoner's dilemma, Cambridge, p. 219–242.

Mäki, U. (2009): Realistic Realism about Unrealistic Models, in: Kincaid, H./Ross, D. [Ed.] (2009): The Oxford handbook of philosophy of economics, Oxford.

Northcott, R./Alexandrova, A. (2015): Prisoner's Dilemma doesn't explain much, in: Peterson, M. [Ed.] (2015): The prisoner's dilemma, Cambridge, p. 64–84.

Ostrom, E. (2009): A Polycentric Approach For Coping With Climate Change.

Patt, A. (2017): Beyond the tragedy of the commons: Reframing effective climate change governance, in: Energy Research & Social Science, Vol. 34, p. 1–3.

Peterson, M. [Ed.] (2015): The prisoner's dilemma, Cambridge.

Poteete, A.R./Ostrom, E./Janssen, M. (2010): Working together. Collective action, the commons, and multiple methods in practice, Princeton, N.J.

Quine, W. (1951): Main Trends in Recent Philosophy: Two Dogmas of Empiricism, in: The Philosophical Review, Vol. 60, No. 1, p. 20.

Stanford, K. (2017): Underdetermination of Scientific Theory, in: Zalta, E. N. [Ed.] (2017): The Stanford Encyclopedia of Philosophy, 2017th. edition.

Stern, N. (2007): The economics of climate change. The Stern review, Cambridge.

Tilman, A.R./Dixit, A.K./Levin, S.A. (2019): Localized prosocial preferences, public goods, and common-pool resources, in: Proceedings of the National Academy of Sciences of the United States of America, Vol. 116, No. 12, p. 5305–5310.

Tversky, A./Kahneman, D. (1973): Availability: A heuristic for judging frequency and probability, in: Cognitive Psychology, Vol. 5, No. 2, p. 207–232.

Veit, W. (2020): Model Pluralism, in: Philosophy of the Social Sciences, Vol. 50, No. 2, p. 91–114.

Veit, W. (2021): Model diversity and the embarrassment of riches, in: Journal of Economic Methodology, p. 1–13.

Ward, J.D./Sutton, P.C./Werner, A.D./Costanza, R./Mohr, S.H./Simmons, C.T. (2016): Is Decoupling GDP Growth from Environmental Impact Possible?, in: PLoS ONE, Vol. 11, No. 10, e0164733.

Weisberg, M. (2015): Simulation and similarity. Using models to understand the world, Oxford.

Weitzman, M.L. (2007): A Review of the Stern Review on the Economics of Climate Change, in: Journal of Economic Literature, Vol. 45, No. 3, p. 703–724.

Wood, P.J. (2011): Climate change and game theory, in: Annals of the New York Academy of Sciences, Vol. 1219, p. 153–170.

Zalta, E.N. [Ed.] (2017): The Stanford Encyclopedia of Philosophy, 2017th. edition.