International Climate Agreements and the Scream of Greta*

Giovanni Maggi† Robert W. Staiger‡

November 2022

Abstract

The world appears to be facing imminent peril, as countries are not doing enough to keep the Earth's temperature from rising to catastrophic levels and various attempts at international cooperation have failed. Why is this problem so intractable? Can we expect an 11^{th}-hour solution? Will some countries, or even all, succumb on the equilibrium path? We address these questions through a model that features the possibility of climate catastrophe and emphasizes the role of international externalities that a country's policies exert on other countries and intertemporal externalities that current generations exert on future generations. Within this setting, we explore the extent to which international agreements can mitigate the problem of climate change. Our analysis illuminates the role that international climate agreements can be expected to play in addressing climate change, and it points to important limitations on what such agreements can achieve, even under the best of circumstances.

*We thank for helpful comments and discussions Scott Barrett, Jonathan Bendor, Klaus Demset, Allan Hsiao, Emanuel Ornelas, Steve Redding, Matthew Turner, participants at the 2021 NBER Future of Globalization conference, the 2021 International Trade and Environmental Policy workshop, the 2022 Annual Meetings of the ASSA, the 2022 IEFS China Annual Conference, the 2022 Political Economy Sustainability Conference, and seminar participants at Bocconi University, Boston College, Florida State University, the Harvard/MIT joint seminar, Princeton University, the Nuremberg Research Seminar in Economics, Singapore Management University, Syracuse University, the University of Oslo and Yale University (Economics and the Leitner Program). Winston Chen, Wei Xiang and Yan Yan provided outstanding research assistance.

†Department of Economics, Yale University; Graduate School of Economics, FGV-EPGE; and NBER.

‡Department of Economics, Dartmouth College; and NBER.

> "Many perceive global warming as a sort of moral and economic debt, accumulated since the beginning of the Industrial Revolution and now come due after several centuries. In fact, ... [t]he story of the industrial world's kamikaze mission is the story of a single lifetime – the planet brought from seeming stability to the brink of catastrophe in the years between a baptism or bar mitzvah and a funeral."

– from David Wallace-Wells, **The Uninhabitable Earth**, 2019 page 4.

1. Introduction

The world appears to be facing imminent peril from climate change. According to the Intergovernmental Panel on Climate Change (IPCC), the costs of climate change will begin to rise to catastrophic levels if warming is allowed to surpass 1.5 degrees Celsius, and countries are not doing enough to keep the Earth's temperature from rising beyond this level: by many accounts the world is on track to warm by almost 3 degrees Celsius by the end of the century.[1] Yet according to one estimate (Jenkins, 2014), most Americans would be unwilling to pay more than $200 a year in support of energy-conserving policies, an amount that is "woefully short of the investment required to keep warming under catastrophic rates" (Zaki, 2019).[2] And various attempts at international cooperation, such as the Kyoto Protocol and the Paris Agreement on Climate Change, have also fallen short. Why is this problem so intractable? Can we expect an 11^{th}-hour solution? Will some countries, or even all, succumb on the equilibrium path?

In this paper we address these questions through a formal model that features the possibility of climate catastrophe and emphasizes two critical issues with which efforts to address climate change must contend: the international externalities that a country's policies exert on other countries, and the intertemporal externalities that current generations exert on future generations. We explore the problems that arise when countries act noncooperatively in this setting, and the extent to which international climate agreements can mitigate these problems.

Previous research has highlighted two challenges that a climate agreement must meet, relating to country participation and enforcement.[3] In this paper we abstract from these well-studied

[1] See, for example, the assessment by Climate Action Tracker at https://climateactiontracker.org/.

[2] And arguably, the policies chosen by U.S. administrations have fallen short of even this low level of the willingness of Americans to pay for such policies.

[3] See for example Barrett, 1994, Harstad, 2012, Nordhaus, 2015, Battaglini and Harstad, 2016 and Harstad, forthcoming on the former, and Maggi, 2016 and Barrett and Dannenberg, 2018 on the latter.

challenges, and focus instead on a limitation that has not been emphasized in the formal literature on climate agreements. This limitation arises from the fact that it is not possible for a climate agreement to include *future* generations in the bargain alongside current generations. Hence, while a climate agreement can in principle address the "horizontal" externalities that arise from the international aspects of emissions choices, it cannot address the "vertical" externalities exerted by a generation's emissions choices on future generations, nor can it address the "diagonal" externalities exerted by a country's current climate policy on future generations in other countries. A key objective of our analysis is to examine the consequences of this limitation of climate agreements in a world where catastrophic outcomes are possible.

We work with a model world economy in which the successive generations of each country make their consumption decisions either unilaterally or within the context of an international climate agreement (ICA), and where utility is derived from consumption and from the quality of the environment. These two dimensions of utility are in tension, as consumption generates carbon emissions, which add to the global carbon stock and degrade the quality of the environment through a warming climate. This tension defines the fundamental tradeoff faced by each generation. In our core analysis we focus on a world without intergenerational altruism, and then later introduce the possibility that each generation cares about its offspring.

When born, a generation inherits the global carbon stock that was determined by the cumulative consumption decisions of the previous generations. As the carbon stock rises, the climate warms and the utility derived by the current generation from the quality of the environment falls commensurately, at least for moderate levels of warming. But if the carbon stock gets too high, the implications are catastrophic: the generation alive at the brink faces the prospect that life could go from livable to essentially unlivable in their lifetime.

We consider two possible scenarios for climate catastrophes. In our common-brink scenario, all countries are brought to the brink of climate catastrophe at the same moment, when the global carbon stock reaches a critical level. In our heterogeneous-brink scenario, more vulnerable countries reach the brink first. As we demonstrate, these two possibilities carry starkly different implications for outcomes along the equilibrium path and for the potential role of an ICA.

We begin with an analysis of the common-brink setting. In the absence of an ICA, we show that the equilibrium path in this setting exhibits an initial warming phase, during which each country's emissions are constant at a "Business-As-Usual" (BAU) level. During this phase, the climate externalities imposed by the emissions choices of a given generation in a given country

on all other countries and on future generations everywhere are left unaddressed, the global stock of carbon rises suboptimally fast, and the implied degradation of the environment erodes the utility of each successive generation, until the world is brought to the brink of catastrophe. Once the brink is reached, however, the brink generation can overcome all of these externalities and avert catastrophe with an 11^{th} hour solution that has each country doing its part to halt further climate change; we show that this 11^{th} hour solution will prevail in any equilibrium in undominated strategies. The solution involves reduced worldwide emissions levels that keep the carbon stock constant given the natural rate of atmospheric regeneration and remain at that level for all generations thereafter, and it implies a precipitous drop in utility for the brink generation and all future generations. The reason for this 11^{th} hour noncooperative solution is that, while earlier generations face rising costs of global warming as their emissions contribute to a growing global stock of carbon, it is only the brink generation that faces the catastrophic implications of continuing the emissions practices of the past. And in the face of this clear and present danger, the nature of the game is fundamentally altered, with the result that the brink generation "does whatever it takes" in the noncooperative equilibrium to avoid catastrophe.

The noncooperative equilibrium in our common-brink scenario therefore delivers a good news/bad news message: the good news is that, while it takes a crisis to shake the world from business-as-usual behavior, when the crisis arrives the world will find a way to save itself from going over the brink; the bad news is that the world that is saved on the brink is not likely to be a nice world in which to live, both because the climate at the brink of catastrophe may be very unpleasant, and because the generation that comes of age at the brink and all future generations must accept a potentially large drop in consumption and utility relative to previous generations in order to prevent the climate from worsening further and resulting in global annihilation. Hence, the brink generation, once born, has an especially strong reason to regret that previous generations did not do more to address climate change.

We next ask: What is the role for an ICA to improve over the noncooperative equilibrium in our common-brink setting? We find that up until the world reaches the brink, the ICA plays the standard role of internalizing the international climate externalities in our model and thereby solves a Prisoner's Dilemma, with all of the participation and enforcement issues that such cooperation involves. But when the world reaches the brink, the role of an ICA is transformed. At most the ICA can solve a coordination problem for its member countries if in the noncooperative equilibrium countries would fail to coordinate on an equilibrium in undominated strategies

and would instead go over the brink, and for this task the need for enforceable commitments over participation and emissions levels disappears; we argue that this may provide a possible interpretation of the evolution from the 1997 Kyoto Protocol, which focused on negotiating and securing enforceable commitments from its members regarding emissions reductions, to the 2016 Paris Agreement, which has moved to an alternative approach to emissions reductions centered on "Nationally Determined Contributions" announced voluntarily by each country. And if coordination can be achieved without an ICA, then according to the common-brink scenario, at the brink of catastrophe a role for the ICA ceases to exist completely.

A remaining question is how the outcome achieved by an ICA compares with the outcome that would be implemented by a global social planner. While the ICA negotiated by each generation in effect picks an extreme point on the Pareto frontier that places zero weight on the utility of future generations directly, we focus on the possibility that the global social planner might instead place strictly positive weight on future generations directly, and hence takes into account not only the horizontal but also the vertical and diagonal climate externalities. We show that, while an ICA slows down the growth of the carbon stock relative to the noncooperative outcome, it does not do so enough relative to the choices of the global social planner. This leads to three possibilities, depending on the severity of the constraint imposed by the catastrophe threshold. If this constraint is sufficiently mild, the ICA will prevent the world from reaching the brink of catastrophe, but the steady-state carbon stock is still too large relative to the social planner outcome; if the constraint is sufficiently severe, the brink will be reached both under the ICA and the social planner, but it is reached at an earlier date under the ICA; and in between these two cases, the world reaches the brink of catastrophe under the ICA but not under the choices of the social planner. It is when the carbon stock constraint lies in this third, intermediate, range that the inability of the ICA to take into account directly the interests of future generations has its most profound impact: while a global social planner would keep the world from ever arriving at the brink of climate catastrophe, an ICA will at best only postpone the arrival at the brink, and when that day arrives, the brink generation and all generations thereafter will suffer a precipitous drop in welfare.

We then turn to the heterogeneous-brink setting, where countries face catastrophe at different levels of the global carbon stock. We assume that if a country were to collapse, its citizens would become climate refugees and suffer a utility cost themselves while also imposing "refugee externality costs" on the remaining countries who receive them. Along the noncooperative path

the world may now pass through three possible phases: a warming phase, where warming takes place but no catastrophes occur; a catastrophe phase, where warming continues and a sequence of countries collapse; and a third phase where warming and catastrophes are brought to a halt. The first and third phases are familiar from the common-brink scenario; the possibility of a middle phase in which some countries collapse along the noncooperative path is novel to the heterogeneous-brink scenario. We show that under mild conditions the world will indeed traverse through all three phases of climate change along the equilibrium noncooperative path – and some of the most vulnerable countries will collapse.

The heterogeneous-brink scenario provides an illuminating counterpoint to our common-brink scenario, where once the world reaches the brink countries do whatever is necessary to avoid global collapse. Relative to that setting, the difference is that each country now has its *own* brink generation, who faces the existential climate crisis *alone* and up against the other countries in the world, who have no reason in the noncooperative equilibrium to internalize the impact of their emissions choices on the fate of the brink country beyond the climate refugee costs that they would incur should the country collapse. And with heterogeneous collapse points, some countries may continue to enjoy a reasonable standard of living once the global carbon stock has stabilized while others have suffered climate collapse, bringing into high relief the potential unevenness of the impacts of climate change across those countries who, due to attributes of geography and/or socioeconomic position, are more or less fortunate. Moreover, even small differences across countries can lead to country collapse along the equilibrium path: unless the brink generation of each country arrives at the same moment, the "we are all in this together" forces that enabled the world to avoid collapse in the noncooperative equilibrium of our common-brink scenario will be disrupted. As we demonstrate, climate refugee externalities will bring back an element of these forces, albeit only partially.

We then revisit the potential role for ICAs, but now in the setting where the catastrophe point differs across countries. We find that the ICA may or may not save a country that would collapse in the noncooperative scenario, but no country will be allowed to collapse under the ICA that would not have collapsed in the absence of the ICA. And we find that, as a result of its inability to take into account directly the interests of future generations, the ICA may allow a range of the most vulnerable countries to collapse when a global social planner would not allow this to happen. We show that these findings hold even in a world where countries can make unlimited international transfers, and that when such transfers are limited by a country's

resource constraints the prospects for small, vulnerable countries such as the Maldives becomes even more bleak. Nevertheless, we also identify a surprising possibility: the outcome desired by a social planner may deviate in the other direction from the ICA outcome, allowing a country to collapse that the ICA would save. This is possible because, while the social planner adopts a carbon policy that redistributes welfare toward future generations relative to the carbon policy of the ICA, the best way to effect this redistribution may be to make an earlier generation face the costs of a country's collapse if that frees future generations from having to live within the constraints imposed by that country's brink.

Finally, we extend our analysis to allow for the presence of intergenerational altruism, and argue that our main qualitative insights are robust to this extension. We also identify some new strategic effects that help shape the noncooperative path of carbon policy when current generations care about the utility of future generations, including a novel "dynamic free-rider effect." In the common-brink scenario, if the world is expected to reach the brink tomorrow, so that all countries will share the burden of avoiding catastrophe, an individual country today has less incentive to keep its emissions low.[4] This effect can exacerbate the overly high level of noncooperative emissions, and may introduce an additional role for an ICA. Furthermore, moving back in time, the anticipation of possible dynamic free-riding behavior can induce countries to keep their emissions below the BAU level, precisely to prevent dynamic free-riding from arising in equilibrium. Interestingly, in this case the possibility of catastrophe affects equilibrium emissions even though the brink is not reached in equilibrium. We also highlight that, in the heterogeneous-brink scenario, an asymmetric version of the dynamic free-rider effect can arise, in that more resilient countries free-ride on the future efforts of less resilient countries to save themselves from collapse.

Overall, our conclusions are sobering. Even abstracting from issues of free-riding in participation and compliance, our model suggests that ICAs can play only a limited role in addressing the most pressing challenges of global warming. If countries face a common threshold of catastrophe, the ICA has a potential role to play during the warming phase, by internalizing the horizontal climate externalities and slowing the world's march to the brink, but it falls short of achieving the outcome of a global social planner, which requires the world to move even

[4] A dynamic free-rider effect arises also in Battaglini et al. (2014, 2016). In their model, higher investment by a player today induces lower investment by other players in the future. In our paper we identify a novel type of dynamic free-rider effect, which is specifically linked to the possibility of catastrophe, a possibility that is absent in the above-mentioned papers.

more slowly toward the brink and possibly avoid the brink altogether. And the ICA has no role to play in saving the world from collapse, beyond a possible role in solving a coordination problem for its member countries, because once the brink of catastrophe is reached countries have sufficient incentives to coordinate on an equilibrium that avoids catastrophe even without an ICA. If the catastrophe threshold varies across countries, the role of an ICA is potentially more expansive, because it may save some of the most vulnerable countries from collapse, but its limitations relative to the global social planner are potentially more devastating, because it may not save enough countries from collapse.

Relative to the existing literature on ICAs, the main contribution of our paper is to analyze the joint implications of international and intergenerational externalities in a world with the potential for catastrophic effects of climate change. We are not aware of any formal analysis that considers the interaction between these fundamental ingredients.

There is an emerging literature that considers optimal environmental policy in the face of climate catastrophe. Prominent examples include Barrett (2013), Lemoine and Rudik (2017) and Besley and Dixit (2019).[5] Of these papers, only Barrett (2013) considers the role of ICAs, but his model is effectively static and does not consider intergenerational issues that we emphasize here. A key point in his paper is that, if the level of the carbon stock that triggers a catastrophe is known with certainty, there exists a noncooperative equilibrium in which no catastrophe occurs, and hence the only possible role for an ICA is to help countries coordinate on the "good" equilibrium – a point that is consistent with our common-brink scenario.[6]

The papers of John and Pecchenino (1997) and Kotlikoff et al. (2021a,b) are also related. Like ours, theses papers consider both international and intergenerational environmental externalities, but they do not consider the possibility of catastrophes. Instead, the central message of John and Pecchenino (1997) is that cooperation between countries at a point in time may be harmful to future generations. This is because there are two international externalities in their model: one stemming from cross-border pollution, and one related to environment-enhancing investments. Internalizing the pollution externality benefits future generations (an effect that is

[5] See also Brander and Taylor (1998), who consider catastrophes in a model that links population dynamics with renewable resource dynamics but does not feature intergenerational or international externalities.

[6] Barrett (2013) also argues that if the catastrophic threshold is uncertain, there is a unique Nash equilibrium that can lead to catastrophe, and an ICA can achieve a Pareto improvement over that equilibrium and reduce the probability of catastrophe. He also emphasizes that, while in the absence of uncertainty the only role of an ICA is to help countries coordinate on the efficient equilibrium without catastrophe, in the setting with uncertainty the ICA has to overcome enforcement and participation issues, just as in models without catastrophes. We discuss how our results can incorporate uncertainty over the brink point in the Conclusion.

present also in our model), but international cooperation on the investment dimension increases the efficiency of resource allocation and hence increases consumption, which tends to degrade the environment. Kotlikoff et al. (2021a,b) focus on characterizing Pareto-improving carbon taxes; they are not concerned with understanding the commitments that could be negotiated in an ICA to address these environmental externalities, which is our central focus here.

Our paper is also related to the literature on the dynamics of ICAs, which includes Dutta and Radner (2004), Harstad (2012, 2021, forthcoming) and Battaglini and Harstad (2016). These papers focus on aspects of ICAs that are very different from the ones we emphasize in this paper, and they do not consider issues of intergenerational externalities or the possibility of catastrophes. In particular, Harstad (2012) and Battaglini and Harstad (2016) focus on issues of free-riding and participation in ICAs when countries can make irreversible investments in green technology that cannot be contracted upon, and Harstad (forthcoming) takes this approach one step further by considering the implications of alternative bargaining procedures.[7] Finally, Harstad (2021) focuses on the desirability of issue linkage through a trade agreement whose commitments are made contingent on forest conservation measures.

The remainder of the paper proceeds as follows. After laying out the basic elements of our modeling framework in section 2, section 3 sets out our common-brink scenario and characterizes the noncooperative emissions choices, as well as those under an ICA and the choices of a global social planner. Section 4 contains the parallel analysis for our heterogeneous-brink scenario. We introduce intergenerational altruism in section 5. Finally, section 6 concludes by discussing a number of further extensions to our core models. An Appendix provides proofs not contained in the body of the paper.

2. The Basic Modeling Framework

We begin by laying out the basic elements of our modeling framework. These elements will form an "umbrella" model, from which our common-brink and heterogeneous-brink scenarios then emerge as special cases.

We consider a world of M countries. Each country has an identical population of identical citizens; we normalize the (initial) population of each country to one. Time is discrete and indexed by $t \in \{0, 1, ..., \infty\}$. We adopt a "successive generations" setting (see Fahri and Wern-

[7] For earlier analyses of ICAs that focus on issues of participation and enforcement, see for example Barrett (1994), Carraro and Siniscalco (1993) and Kolstad and Toman (2005).

ing, 2007), where the citizen in each country lives for one period and is replaced by a single descendant in the next period. We allow each parent to be altruistic toward its only child, and the per-capita utility of country i's generation t is given by

$$\tilde{u}_{i,t} = u_{i,t} + \beta \tilde{u}_{i,t+1}$$

where $u_{i,t}$ is material per-capita utility and the parameter $\beta \geq 0$ captures the degree of intergenerational altruism. In this setting, utility can be equivalently represented with the dynastic utility function

$$\tilde{u}_{i,t} = \sum_{s=0}^{\infty} \beta^s u_{i,t+s}. \tag{2.1}$$

Material utility $u_{i,t}$ is derived from consumption and from the quality of the environment. But these two dimensions of utility are in tension, as consumption generates carbon emissions, which add to the global carbon stock and degrade the quality of the environment through a warming climate. This tension defines the fundamental tradeoff faced by each generation.

To highlight this tradeoff, we abstract from trading relations between countries, so that we can focus on their interactions mediated through the global carbon stock.[8] And we adopt a reduced form approach to modeling the consumption benefits of emissions, by specifying the benefits directly as a function of emissions rather than the underlying consumption choices that generate the emissions. In particular, we use the increasing and concave function $B(c_{i,t})$ to denote these benefits, where $c_{i,t} \geq 0$ is the level of carbon emissions of country i's generation t.[9] We therefore treat $c_{i,t}$ itself as the choice variable of country i, with the understanding that lower emissions mean lower consumption.[10] We have in mind that each government i then implements its chosen $c_{i,t}$ with an appropriate climate policy (e.g., carbon tax).

[8] We briefly consider the implications of trade in the Conclusion.

[9] Our restriction that $c_{i,t} \geq 0$ reflects the possibility of zero (net) emissions through carbon capture and other mitigation efforts. By this logic we could impose $c_{i,t} \geq c^{\min}$ where $c^{\min}$ could be strictly positive or even strictly negative, but in our formal analysis it is convenient to abstract from these possibilities and equate the emissions generated by a country's best mitigation efforts with its emissions were it to collapse.

[10] Implicit in our specification of the reduced-form benefit function $B(c_{i,t})$ is the assumption that there is a one-to-one mapping between a country's emissions and its utility from consumption, and hence that the stock of carbon does not itself impact this mapping (e.g., by impacting a country's productivity associated with any level of emissions). In principle we could allow for such an impact with an alternative benefit function $B(c_{i,t}, C_t)$ where C_t shifts down (or up) the benefit function for a given level of emissions. If it were allowed that $B_{c_{i,t}C_t} \neq 0$ then this would complicate the paths of emissions that we derive below, but we do not believe that it would generate interesting qualitative differences in the results we emphasize, and so we opt for the tractability of our simpler modeling assumption.

While a country's own period-t emissions generate consumption benefits for its generation t, these emissions also contribute to the global stock of carbon in the atmosphere. We denote by C_t the global carbon stock in period t.[11] The evolution of this stock through time depends on the depreciation rate of the stock and on the level of emissions $c_{i,t}$, according to

$$C_t = (1-\rho)C_{t-1} + c_t^W, \text{ with } C_{-1} = 0, \tag{2.2}$$

where c_t^W denotes aggregate world emissions at time t. The parameter $\rho \in [0,1)$ reflects the natural rate of atmospheric "regeneration": if $\rho = 1$, by the beginning of the current period the previous period's stock of carbon is gone; if $\rho = 0$, the current period inherits the full stock of carbon from the previous period. As will become clear below, the relationship in (2.2) implies that each generation feels the impact of its own emissions (because these emissions add to the carbon stock in the current period).[12]

We assume that increases in the global carbon stock degrade the environment and lead to losses in material welfare. In particular, we assume for country i that these losses rise linearly with the global carbon stock C_t according to the parameter $\lambda > 0$ as long as C_t is below country i's threshold level $\tilde{C}_i$; but if C_t exceeds this threshold, country i collapses and its citizens become climate refugees, suffering a one-time per-capita material utility cost $L > 0$. Moreover, we assume that a collapsing country's refugees spread uniformly across the remaining countries, with each refugee imposing a one-time material utility cost r on the country to which it immigrates.[13] We have in mind that moderate degrees of global warming lead to moderate costs for a country. But past a certain critical level, a rising carbon stock leads to a level of global warming that would be catastrophic for the country, triggering its collapse. The catastrophic level $\tilde{C}_i$ is assumed known with certainty.[14] And to focus on the main points, we assume that, aside from these critical threshold levels, countries are symmetric in all respects.

Our modeling framework highlights two externalities that arise in the context of climate change. One externality is international: the emissions of one country's generation t contribute

[11] More accurately, C_t can be thought of as the atmospheric CO_2 concentration, an increase in which leads to global warming. But in the text we will refer to C_t as the carbon stock.

[12] Given that a period corresponds to a generation in our model, this feature seems broadly realistic, as existing estimates put the time it takes for current carbon emissions to translate into higher global temperatures at between 10 and 40 years (see, for example, Pindyck, 2020, who also reports an estimate of the dissipation rate ρ on the order of 0.0035 per year).

[13] Another potential cost imposed by the collapse of a country on the surviving countries would be the destruction of international trade between these countries, a possibility we return to briefly in the Conclusion.

[14] We discuss the more realistic possibility that $\tilde{C}_i$ is uncertain in the Conclusion.

to the global stock of period-t carbon, which impacts the material utility of generation-t in all other countries. The other externality is intergenerational: the emissions of a country's generation t affect the material utility of all subsequent generations $t+1$ and beyond in that country. Moreover, these "horizontal" (international) and "vertical" (intergenerational) externalities interact to produce additional "diagonal" externalities: the emissions of one country's generation t impact the utility of future generations in all other countries.

In the sections to follow we will characterize three regimes where these externalities are addressed to varying degrees. In the noncooperative equilibrium that arises if countries choose emissions levels in the absence of any agreements, neither the horizontal nor the vertical or diagonal externalities are addressed, in the sense that each country's emissions choices impose costs on other countries and on future generations that those parties did not agree to incur. In the ICA equilibrium that arises when international agreements over emissions levels are possible and international lump-sum transfers are available, we argue below that only the horizontal externalities can be addressed, not the vertical or diagonal externalities: as future generations cannot sit at the table while the ICA is negotiated, even in the presence of intergenerational altruism the ICA's emissions choices inevitably impose costs on future generations in all countries that those parties did not agree to incur. And third, we consider as a benchmark the social optimum that a global "social planner" would implement beginning at time $t = 0$ to maximize a social welfare function. We assume that the social welfare function puts positive weights on future generations *directly*, not just indirectly through the intergenerational altruism of the initial generation; and we assume that intergenerational lump-sum transfers are unavailable, leaving emissions and international lump-sum transfers as the planner's choice variables.[15]

More specifically, to characterize the global social planner's choices we follow Fahri and Werning (2007) in postulating the following planner objective:[16]

$$W = \sum_{t=0}^{\infty} \hat{\beta}^t \tilde{U}_t \tag{2.3}$$

[15] An alternative benchmark would be the "unconstrained first best," meaning that all generations and countries can strike a Coasian bargain where also intergenerational transfers are available. As we will argue below (see note 18), the key difference between the unconstrained first best and our social optimum amounts to the weights that are placed on the utility of future generations when choosing emissions. We choose not to focus on the unconstrained first best because intergenerational transfers are arguably not feasible in reality, as discussed below. On the other hand, our social optimum can be interpreted as the policies that would be chosen by an ICA were there to be a shift in power toward younger generations, a benchmark that seems more relevant to the current debate on climate change.

[16] See also Caplin and Leahy (2004), Feng and Ke (2018), and Millner and Heal (2021).

where $\tilde{U}_t$ is average per-capita world-wide utility from the point of view of generation t and $\hat{\beta}$ is the planner's discount factor. Notice that regardless of the degree of intergenerational altruism displayed by each generation, there will be a discrete wedge between the social and private discount factor $(\hat{\beta} - \beta)$ as long as the planner puts strictly positive weights on future generations directly, hence we have $\hat{\beta} > \beta$. Moreover, in general this wedge need not decrease as β rises. For example, in a two-period setting with α the Pareto weight placed by the planner on the second generation, we would have $\hat{\beta} = \beta + \alpha$. Notice also that in principle $\hat{\beta}$ could be greater than one, but to avoid the complications that would arise if this were the case we assume for simplicity that $\hat{\beta} < 1$.[17]

It is worth pausing to clarify the nature of the discrepancy between the social planner's choice and that of the ICA. To this end, suppose for a moment that there is no intergenerational altruism ($\beta = 0$). In this case, when generation t chooses emissions to maximize its utility under the ICA, it ignores the impact of these emissions on future generations and simply maximizes its own material utility. A planner who puts positive weight on each generation ($\hat{\beta} > 0$) would modify the choices of generation t and redistribute utility from generation t to subsequent generations; the same logic applies also in the presence of altruism ($\beta > 0$), because as noted above, if the social welfare function puts direct weight on future generations the wedge between the social and private discount factors $(\hat{\beta} - \beta)$ need not decrease as β rises. Notice, though, that the social planner's choice does not mark a Pareto improvement over the ICA, but rather a movement along the efficiency frontier, shifting surplus from generation t to later generations.

Finally, before turning to the analysis we can make a simple preliminary point: the impacts associated with horizontal and vertical externalities *reinforce* each other. This can be seen most clearly by focusing on a special and simple case of our model, in which there is no catastrophe point ($\tilde{C}_i = \infty$ for all i) and no intergenerational altruism ($\beta = 0$); and by comparing the noncooperative emissions choices to those chosen by the social planner. In this case it is straightforward to show and intuitive that in the noncooperative equilibrium each country's generation t would choose a level of emissions to satisfy $B'(c_t) = \lambda$ (assuming interior solutions), while the emissions levels chosen by the planner for each country's generation t satisfy $B'(c_t) = \frac{M}{1-\hat{\beta}(1-\rho)}\lambda$. The overall wedge between the planner's emissions choices and noncooperative emissions choices is summarized by $\frac{M}{1-\hat{\beta}(1-\rho)} > 1$, which implies excessive emissions in

[17] In the case where $\hat{\beta} \geq 1$, the infinite sum in (2.3) does not converge, so we would have to assume a finite horizon.

the noncooperative equilibrium relative to the planner's outcome. The wedge has two components: $M > 1$ reflects the degree to which the international externality contributes to excessive emissions, because noncooperative choices do not account appropriately for the environmental costs of a country's emissions that are imposed on other countries; and $\frac{1}{1-\hat{\beta}(1-\rho)} > 1$ reflects the degree to which the intergenerational externality contributes to excessive emissions, because noncooperative choices do not account for the environmental costs of a country's emissions that are imposed on future generations. The two externalities enter multiplicatively into this wedge, so they reinforce each other. Intuitively, this is a consequence of the above-mentioned fact that there are not only "horizontal" and "vertical" externalities, but also "diagonal" externalities.[18]

This special case is useful for highlighting in simple terms the impacts of the externalities that arise in our model. But it is also useful as a benchmark to illustrate the critical role that the existence of a catastrophe point ($\tilde{C}_i$ finite for at least some i) plays in our analysis of climate policy. In the absence of a catastrophe point, the social planner and noncooperative emissions profiles are straightforward, as we have just observed, as is the emissions profile under an ICA. But as we establish below, the existence of a catastrophe point introduces fundamental changes to these emissions profiles, both along the path to the catastrophe and once the brink of catastrophe is reached, and it changes the possible role of an ICA as well.

The importance of a catastrophe point for understanding the policy challenges posed by climate change is one of the central messages of our paper. To deliver this message, we henceforth focus on the case in which $\tilde{C}_i$ is finite for at least some i. We will proceed by focusing for now on a world without intergenerational altruism ($\beta = 0$); in section 5 we consider as well the possibility that $\beta > 0$ and show how our results extend in the presence of intergenerational altruism. Notice from (2.1) that with $\beta = 0$ there is no distinction between utility ($\tilde{u}_{i,t}$) and material utility ($u_{i,t}$), and for this reason in what follows we will simply refer to "utility" and use the notation $u_{i,t}$ to denote the utility of country i's generation t (and similarly the notation U_t to denote average per-capita world-wide utility of generation t).

[18] In this setting without the possibility of catastrophe, it is easy to see the difference between our social optimum and the unconstrained first best (as defined in note 15). If all countries and generations could strike a Coasian bargain and intergenerational transfers were available (and assuming transferrable utility), the emissions level would maximize the sum of utilities of all parties to the bargain, and therefore would satisfy $B'(c) = \frac{M}{\rho}\lambda$. Intuitively, the Coasian solution takes into account the externality that emissions impose on all present and future citizens in equal measure. Note also that, given our assumption $\hat{\beta} < 1$, the unconstrained first best would go further than our social optimum in cutting emissions, and so in this sense the choices of our social planner are more conservative relative to the unconstrained first best outcome.

3. The Common-Brink Scenario

We first consider a world described by the basic modeling framework laid out in section 2, but where $\tilde{C}_i = \tilde{C}$ for all i, so that all countries share a common level of the carbon stock that would bring them to the brink of catastrophe. We will refer to this as the common-brink scenario.

In this world, countries are fully symmetric, so we can omit the country subscript i. Here, moderate degrees of global warming lead to moderate costs, but past a certain critical level a rising carbon stock leads to a level of global warming that would trigger the collapse of civilization.[19] Since in this world climate refugees have nowhere to go, we suppose L is extremely high ($L = \infty$). The utility of a representative citizen in generation t is then given by

$$u_t = \begin{cases} B(c_t) - \lambda C_t & \text{if } C_t \leq \tilde{C} \\ -\infty & \text{if } C_t > \tilde{C}. \end{cases} \tag{3.1}$$

3.1. Noncooperative Equilibrium

We begin our analysis of the common-brink scenario by characterizing the noncooperative emissions choices. As we noted above, we assume $\beta = 0$ for now, so that there is no intergenerational altruism. Given $\beta = 0$, countries are effectively myopic and we can focus on (subgame perfect) Nash equilibria. This implies that the noncooperative equilibrium in general has two phases.

The first phase is a "warming phase," during which the emissions of each country's generation t is constant at the level $\bar{c}^N$ defined by $B'(\bar{c}^N) = \lambda$, where the marginal benefit to each country of the last unit of carbon that it emits is equal to the marginal loss of utility that it suffers as this unit of carbon is added to the global carbon stock, implying

$$\bar{c}^N = B'^{-1}(\lambda). \tag{3.2}$$

[19] In reality, collapse on a global scale is unlikely to be the result of crossing a single climate threshold. Rather, as Wallace-Wells (2019) argues forcefully, it is the cumulative effect of the collapse of numerous ecological subsystems, each cascading over their own "tipping points," that poses the most serious climate-change induced existential threat to civilization (see Lenton et al., 2008, for an attempt to identify the location of tipping points for a variety of ecological subsystems). While therefore highly stylized along this dimension, our common-brink scenario might nevertheless be viewed as approximating a world in which there are many intermediate thresholds for the carbon stock that define a step function for the cost of global warming that over an initial range is composed of many small steps (approximated in our model by a smooth, linearly increasing cost), and where the brink as we have defined it is then associated with a carbon threshold level that, if crossed, would be the tipping point for a final ecological subsystem that would prove to be the "straw that broke the camel's back." The defining feature of the common-brink scenario is that all of the countries of the world arrive at the brink together. We postpone until the next section consideration of the possibility that each country could face its own brink level of the global carbon stock, and hence that some countries may be more vulnerable to the effects of climate change than others.

As is intuitive, (3.2) implies that $\bar{c}^N$ is decreasing in λ, the marginal cost in terms of own utility associated with another unit of carbon emissions. We can think of $\bar{c}^N$ as corresponding to "Business-As-Usual" (BAU) emissions levels. During the warming phase associated with these choices, the global stock of carbon grows according to

$$C_t^N = (1-\rho)C_{t-1}^N + M\bar{c}^N, \text{ with } C_{-1}^N = 0, \tag{3.3}$$

and as the global carbon stock C_t^N grows and the cost of climate change mounts, the utility of each successive generation in every country declines according to

$$u_t^N = B(\bar{c}^N) - \lambda C_t^N. \tag{3.4}$$

If the warming phase went on forever, (3.3) implies that the global carbon stock would converge to the steady state level $\frac{M}{\rho}B'^{-1}(\lambda) \equiv C^N$. And if the catastrophe level of the global carbon stock, $\tilde{C}$, were greater than C^N, then BAU emissions could indeed go on forever without triggering a climate catastrophe. But the view of the majority of climate scientists is that a climate catastrophe will occur in finite time, perhaps by the end of this century, if the world stays on a BAU emissions path (see the recent reports of the IPCC). In the language of our model this view translates into a statement that $\tilde{C}$ lies below C^N. We therefore impose

$$\tilde{C} < C^N \qquad \text{(Assumption 1)}$$

which ensures that under BAU emissions the catastrophic level of the global climate stock would eventually be breached.

The second phase of the noncooperative equilibrium kicks in when C_t^N reaches the brink of catastrophe $\tilde{C}$. This occurs for the "brink generation" $t = \tilde{t}^N$ where, ignoring integer constraints, $\tilde{t}^N$ is defined using (3.3) by $C_{\tilde{t}^N}^N = \tilde{C}$. In effect, $\tilde{t}^N$ represents the point in time where, in a single generation, life under BAU emissions would go from livable to unlivable.

If the brink generation $\tilde{t}^N$ is to avoid the collapse of civilization, it must end the warming phase with an "11^{th}-hour solution" that brings climate change to a halt. Indeed it is easy to see that if it is feasible to do so, then at any equilibrium in undominated strategies, C_t^N remains at $\tilde{C}$ for $t = \tilde{t}^N$ and also for all subsequent generations. Focussing on the symmetric equilibrium in undominated strategies, for generations $t \geq \tilde{t}^N$ emissions will fall to the replacement level dictated by the natural rate of atmospheric regeneration given by

$$c_t = \frac{\rho\tilde{C}}{M} \equiv \hat{c}^N \tag{3.5}$$

where $\hat{c}^N < \bar{c}^N$ is implied by Assumption 1. With $c_t = \hat{c}^N$ for generations $t \geq \tilde{t}^N$, the world remains on – but does not go over – the brink of catastrophe, so the collapse of civilization is avoided. To confirm that $\hat{c}^N$ is indeed an equilibrium emissions level for generations $t \geq \tilde{t}^N$, we need only note that unilateral deviation to an emissions level higher than $\hat{c}^N$ would trigger climate catastrophe and infinite loss, while deviation to a lower emissions level would not be desirable either given that $\hat{c}^N < \bar{c}^N$.

While here we emphasize symmetric equilibria in undominated strategies, it should be noted that there exist two other types of equilibria. First, there are equilibria where the world collapses, because if other countries choose very high emission levels, an individual country is indifferent over its own emission levels, so it is an equilibrium for all countries to choose very high emission levels. But it is easy to see that such equilibria are in weakly dominated strategies: starting from such an equilibrium, a country can weakly improve its payoff by lowering its emissions. And second, there is a continuum of asymmetric equilibria (in undominated strategies) where the world survives, with some countries cutting their emissions levels below $\hat{c}^N$ while others raise their emissions levels above $\hat{c}^N$ and the sum of world-wide emissions remains at the level $\rho\tilde{C}$ which holds the world at the brink. It is easy to see that these asymmetric equilibria are inefficient given our symmetric-country setup, and so we take the symmetric equilibrium as the natural focal point. As we will discuss below, in the event that countries focus on one of the asymmetric and inefficient Nash equilibria, or even worse, if countries fail to coordinate at all and focus on a catastrophic equilibrium where the brink is crossed, then a coordination role for an international climate agreement would arise.

Returning to our analysis of the symmetric equilibrium in undominated strategies, the utility of each generation $t \geq \tilde{t}^N$ during this second phase of the noncooperative equilibrium is then constant and given by

$$u_t^N = B(\hat{c}^N) - \lambda\tilde{C}. \tag{3.6}$$

We may conclude that the noncooperative emissions path for each country is given by

$$c_t^N = \begin{cases} \bar{c}^N & \text{for } t < \tilde{t}^N \\ \hat{c}^N & \text{for } t \geq \tilde{t}^N. \end{cases} \tag{3.7}$$

Combining (3.6) with (3.4) we then also have the path of noncooperative utility:

$$u_t^N = \begin{cases} B(\bar{c}^N) - \lambda C_t^N & \text{for } t < \tilde{t}^N \\ B(\hat{c}^N) - \lambda\tilde{C} & \text{for } t \geq \tilde{t}^N. \end{cases} \tag{3.8}$$

Note that under the noncooperative equilibrium and according to (3.8), utility must fall precipitously for the brink and all subsequent generations.[20] This is due to the drop in global emissions implied by

$$\hat{c}^N = \frac{\rho \tilde{C}}{M} < B'^{-1}(\lambda) = \bar{c}^N \tag{3.9}$$

that is required to prevent catastrophe once the world reaches the brink, where the inequality in (3.9) follows from Assumption 1 as we have noted. According to (3.8) and (3.9), in order to prevent the planet from warming further, the brink generation and all future generations accept the reduced level of consumption associated with the emissions level $\hat{c}^N$. This consumption level is further below the consumption level enjoyed by previous generations and associated with the emissions level $\bar{c}^N$, (i) the greater the number of countries M, (ii) the smaller the regeneration capacity of the atmosphere ρ and level of carbon stock above which climate catastrophe occurs $\tilde{C}$, and (iii) the lower the cost of moderate pre-catastrophe warming λ.

Summarizing, we may now state:

Proposition 1. *The noncooperative equilibrium in the common-brink scenario exhibits an initial warming phase where each country's emissions are constant at a "Business-As-Usual" level. During this phase, the global stock of carbon rises and the world is brought to the brink of catastrophe. Once the brink is reached, a catastrophe is avoided with an 11^{th} hour solution that halts further climate change with reduced emissions that are set at the replacement level dictated by the natural rate of atmospheric regeneration and remain at that level for all generations thereafter, and which imply a precipitous drop in utility for the brink generation and all future generations.*

Notice an interesting feature of the noncooperative equilibrium described in Proposition 1: no generation up until the brink generation does anything to address the climate externalities that each generation is imposing on those of its generation residing in other countries and on future generations everywhere; and yet the brink generation overcomes all of these externalities and saves the world. The reason for this 11^{th} hour noncooperative solution to the threat of global annihilation posed by climate change is that, while earlier generations face rising costs of global warming as their emissions contribute to a growing global stock of carbon, it is only the brink generation that faces the catastrophic implications of continuing the emissions practices of the

[20] Here and throughout we use the adjective "precipitous" to describe a decline that would remain discrete even in the limit as time in our model went from discrete to continuous.

past. And in the face of this potential catastrophe, the nature of the game is fundamentally altered, with the result that the brink generation "does whatever it takes" in the noncooperative equilibrium to avoid catastrophe.[21]

Hence, Proposition 1 describes a good news/bad news feature of the noncooperative equilibrium: the good news is that, while it takes a crisis to shake the world from business-as-usual behavior, when the crisis arrives the world will find a way to save itself from going over the brink; the bad news is that the world that is saved on the brink is not likely to be a nice world in which to live, both because the climate at the brink of catastrophe may be very unpleasant, and because the brink and all future generations must accept a precipitous drop in consumption and utility in order prevent the climate from worsening further and resulting in annihilation.

3.2. International Climate Agreements

We are now ready to consider what an ICA can achieve. Two important challenges that an ICA must meet relate to participation and enforcement. It is well known (see, for example, Barrett, 1994, Harstad, 2012, Nordhaus, 2015, Battaglini and Harstad, 2016 and Harstad, forthcoming) that ICAs create strong incentives for countries to free ride on the agreement, and that without some means of requiring participation the number of countries participating in an ICA is likely to be very small. And even among the willing participants, there is a serious question of how the commitments agreed to in the ICA can be enforced, given that the agreement must ultimately be self-enforcing and that retaliation using climate policy for this purpose is arguably ineffective (see, for example, Maggi, 2016 and Barrett and Dannenberg, 2018 on the possibility of linking trade agreements to climate agreements in this context). Together these challenges are understood to place important limitations on what an ICA can achieve.

Here we abstract from these well-studied (but in principle, not insurmountable) limitations, and assume that the ICA attains full participation of all M countries in the world, that the noncooperative equilibrium is the "threat point" for the negotiations over an ICA, and that any agreement negotiated under the ICA is perfectly enforceable by an external enforcement mechanism. Under these ideal conditions, we ask what an ICA can accomplish. Our answer highlights an additional limitation that has not been emphasized in the formal literature on climate agreements. This limitation arises from the fact that it is not possible for an ICA to

[21] Barrett (2013) makes a related observation. He notes that the nature of the game can change if countries face a catastrophic loss function associated with climate change, but his observation is made within a static model and emphasizes the implications for the self-enforcement constraint in international climate agreements.

include *future* generations in the bargain alongside current generations. Hence, while an ICA can in principle address the horizontal externalities that arise from the international aspects of emissions choices and that create inefficiencies in the noncooperative outcomes, it cannot address the vertical and diagonal externalities that are associated with the intergenerational aspects of the climate problem. Our goal is to characterize what an ICA can achieve in the presence of this particular limitation, relative to the noncooperative outcome and relative to a global social planner who is not subject to this limitation.[22]

Recalling that we are focusing for now on the case where $\beta = 0$ so as to abstract from intergenerational altruism, for each generation t we characterize the ICA emissions levels as those that maximize the welfare of generation t in the representative country. Given our symmetric-country setting, this is the natural ICA design to focus on, as it would emerge if countries bargain efficiently and have symmetric bargaining power.

Using (3.1), it is direct to confirm that, for as long as the catastrophe point $\tilde{C}$ is not hit, emissions levels under the ICA satisfy $B'(c_t) = M\lambda$ and are hence given by

$$\bar{c}^{ICA} = B'^{-1}(M\lambda). \tag{3.10}$$

According to (3.10), in any period where the catastrophe point is not hit, each country's emissions under the ICA will equate that country's marginal utility from a small increase in emissions to the marginal environmental cost, taking into account the costs imposed on the current generation in all M countries. Notice that (3.2) and (3.10) imply $\bar{c}^{ICA} < \bar{c}^{N}$, because under noncooperative choices each country internalizes the costs imposed on the current generation only in its own country. Finally, with emissions levels given by $\bar{c}^{ICA}$, as long as the brink of catastrophe is not hit the carbon stock under the ICA evolves according to

$$C_t^{ICA} = (1-\rho)C_{t-1}^{ICA} + M\bar{c}^{ICA} \quad \text{with } C_{-1}^{ICA} = 0, \tag{3.11}$$

which defines a process of global warming in which the global carbon stock would eventually converge to the steady state level $\frac{M}{\rho}B'^{-1}(M\lambda) \equiv C^{ICA}$.

[22] One could imagine that in principle an ICA might involve an implicit contract of some kind between current and future generations to internalize the intergenerational externalities. But recall that altruism itself cannot address this issue. Rather, for such a contract to be implemented, future generations would have to be able to punish current generations for any deviations from the contract, and current and future generations would need to find a way to coordinate on a particular equilibrium of this kind even though communication between them is impossible. We view these challenges as essentially insurmountable. Notice also that while the Pareto-improving carbon taxes characterized by Kotlikoff et al. (2021a,b) do focus on internalizing the intergenerational externalities associated with climate policy, they do not speak to the issues we address here, because they are not concerned with characterizing the outcome of international negotiations over climate policy.

Recall that under Assumption 1 the brink of climate catastrophe will be reached under the BAU emissions of the noncooperative equilibrium. Will the ICA keep the world from ever reaching the brink? The answer is yes, if and only if

$$\tilde{C} \geq C^{ICA}, \tag{3.12}$$

where note from their definitions that $C^{ICA} < C^N$ so both Assumption 1 and (3.12) will be satisfied if $\tilde{C} \in [C^{ICA}, C^N)$. Intuitively, if the catastrophe point of the global carbon stock, $\tilde{C}$, is high and sufficiently close to the steady state level of the global carbon stock under BAU emissions, C^N, then only a relatively small reduction in emissions from the BAU level would be required to keep the world from reaching the brink, and by addressing the horizontal externalities the ICA will indeed deliver the required reductions; and the threshold level of the carbon stock C^{ICA} in (3.12) defines "sufficiently close" in this context.

On the other hand, if $\tilde{C}$ is below this threshold level and (3.12) is violated so that

$$\tilde{C} < C^{ICA}, \tag{3.13}$$

then under the ICA the brink of catastrophe will be reached in finite time, and the brink generation $\tilde{t}^{ICA}$ is determined from (3.11) as the period $\tilde{t}^{ICA}$ that satisfies $C^{ICA}_{\tilde{t}^{ICA}} = \tilde{C}$. Notice from (3.11) and (3.3) that $\tilde{t}^{ICA} > \tilde{t}^N$ is ensured by $\bar{c}^{ICA} < \bar{c}^N$, so the ICA postpones the arrival of the brink generation when (3.13) is satisfied even though it does not avoid the brink completely in this case.

If (3.13) is satisfied, what happens under the ICA when the world reaches the brink? This might seem to be when the ICA can play its most important role, by ensuring the very survival of civilization. And clearly, given the utility function in (3.1), the ICA will not let the world go over the brink. But recall that neither would countries go over the brink in the noncooperative equilibrium, as long as they can coordinate on a Nash equilibrium in undominated strategies, the case we have emphasized above. In that case, therefore, the ICA becomes redundant for $t \geq \tilde{t}^{ICA}$, because from that point forward the ICA can do no better than to replicate the noncooperative emissions choices $\hat{c}^N$, and hence, $\tilde{t}^{ICA}$ marks the end of the useful life of the ICA. On the other hand, if one admits the possibility that in the noncooperative game countries might fail to coordinate on a Nash equilibrium in undominated strategies, and as a result the world might collapse in equilibrium, then the ICA would have an important coordination role to play, namely, ensuring that countries stay away from catastrophic Nash equilibria where

the carbon stock exceeds the threshold $\tilde{C}$. Hence, the more general message of our analysis is that the role of an ICA changes dramatically once the world reaches the brink of catastrophe: at that point, the role of an ICA will at most be to address coordination failures among its member countries, and it may even become redundant and have no role to play at all.[23]

Finally, letting c_t^{ICA} denote the path of emissions under the ICA, we can describe emissions succinctly under both (3.12) and (3.13) with

$$c_t^{ICA} = \begin{cases} \bar{c}^{ICA} & \text{for } t < \tilde{t}^{ICA} \\ \hat{c}^N & \text{for } t \geq \tilde{t}^{ICA} \end{cases} \tag{3.14}$$

where $\tilde{t}^{ICA}$ is finite if and only if (3.13) is satisfied. Utility under the ICA is then given by

$$u_t^{ICA} = \begin{cases} B(\bar{c}^{ICA}) - \lambda C_t^{ICA} & \text{for } t < \tilde{t}^{ICA} \\ B(\hat{c}^N) - \lambda \tilde{C} & \text{for } t \geq \tilde{t}^{ICA}. \end{cases} \tag{3.15}$$

Note that under the ICA, if (3.13) is satisfied so that $\tilde{t}^{ICA}$ is finite, then (3.15) implies that utility must fall precipitously for the brink and all subsequent generations, due to the reduction in global emissions implied by

$$\hat{c}^N = \frac{\rho \tilde{C}}{M} < B'^{-1}(M\lambda) = \bar{c}^{ICA} \tag{3.16}$$

that is required to prevent catastrophe once the world reaches the brink, where the inequality in (3.16) follows from (3.13). Hence, according to (3.15) and (3.16) and similar to the noncooperative equilibrium, in order to prevent the planet from warming further, under the ICA the brink generation and all future generations accept a reduced level of consumption. However, with $\bar{c}^{ICA} < \bar{c}^N$ it is also clear that the brink generation suffers a smaller decline in welfare under the ICA than in the noncooperative equilibrium.

We summarize with:

[23] There is another possible coordination role that the ICA might play when the world is at the brink of catastrophe, albeit a less dramatic one than preventing a catastrophic coordination failure. Recall that, in the noncooperative game, we have focussed on the symmetric equilibrium in undominated strategies, which implies that for $t \geq \tilde{t}^N$ countries not only avoid catastrophe, but also adopt the efficient assignment of emissions. If instead countries coordinated on an asymmetric equilibrium (in undominated strategies) for $t \geq \tilde{t}^N$, then the ICA would have an efficiency-enhancing role to play for $t \geq \tilde{t}^{ICA}$, allowing countries to exchange emissions cuts for transfers. In particular, countries would agree to the symmetric and efficient Nash emissions levels $\hat{c}^N$ and use international lump-sum transfers to distribute according to bargaining powers the surplus gains that result from eliminating the inefficiency. In this case the ICA would have a continuing role in enhancing the efficiency properties of the emissions cuts required for survival.

Proposition 2. *In the common-brink scenario the path of emissions under the ICA falls into one of two cases. If $\tilde{C}$ is above a threshold level, then the brink of catastrophe is never reached, and the ICA emissions levels are below the noncooperative levels and constant through time. Otherwise, if $\tilde{C}$ is below this threshold, then the brink of catastrophe will be reached, but at a later date than in the absence of the ICA. In this second case, the ICA emissions levels are below the noncooperative levels and constant through time until the brink is reached, at which point emissions fall to the replacement rate dictated by the natural rate of atmospheric regeneration and remain at that level for all generations thereafter; and the path of ICA emissions implies a precipitous drop in utility for the brink generation relative to the previous generation, but this drop is smaller than under the path of noncooperative emissions levels. If under the ICA the brink is reached, the ICA has a role to play in helping the world avoid climate catastrophe only insofar as it may help countries avoid a coordination failure.*

It is interesting to reflect more broadly on the evolving role of an ICA according to our common-brink scenario. Up until the world reaches the brink, the ICA plays the standard role of internalizing the international climate externalities associated with $\lambda > 0$ in our model and thereby solves a Prisoner's Dilemma, with all of the participation and enforcement issues that this cooperation involves. But when the world reaches the brink, the role of an ICA is transformed: at most it can solve a coordination problem for its member countries, and for this task the need for enforceable commitments over participation and emissions levels disappears. In this light it is relevant to observe that while the 1997 Kyoto Protocol focused on negotiating and securing enforceable commitments from its members regarding emissions reductions, the 2016 Paris Agreement has moved to an alternative approach to emissions reductions centered on "Nationally Determined Contributions" announced voluntarily by each country; a possible interpretation of this evolution through the lens of our common-brink scenario is that the world is reaching the brink, and the role of an ICA is evolving from cooperation to coordination.[24] And if coordination can be achieved without an ICA, then according to the common-brink scenario at the brink of catastrophe a role for the ICA ceases to exist completely. This might seem surprising, since an ICA is able to address horizontal externalities, and the possibility of catastrophe does imply extreme horizontal externalities once the world reaches the brink. But at the brink, these extreme international externalities are coupled with extreme *internalized*

[24] See Barstad (forthcoming) for an alternative interpretation of the evolution from Kyoto to Paris that focuses on changes in the numbers of major carbon polluting countries over this period.

costs of increasing emissions, and this makes ICAs redundant as a means to avoid catastrophe *once the catastrophe is at hand*, because at that point countries have sufficient incentives to avoid catastrophe even in the noncooperative scenario.

3.3. The Social Optimum

We next consider the emissions levels that a global social planner would choose in order to maximize world welfare, or equivalently, given our symmetric setting and using (2.3) and (3.1), the welfare of the representative country

$$W = \sum_{t=0}^{\infty} \hat{\beta}^t u_t.$$

We will refer to the choices of the planner as the socially optimal choices.[25]

Clearly, the planner will not allow the world to end in catastrophe and hence will not allow C_t to exceed $\tilde{C}$. Consequently, for the planner's problem and using (3.1) we can equate u_t with $B(c_t) - \lambda C_t$ and introduce the constraint $C_t \leq \tilde{C}$, which we will henceforth refer to as the "brink constraint." To determine the socially optimal emissions choices, we therefore write the planner's problem as

$$\begin{aligned} & \max \sum_{t=0}^{\infty} \hat{\beta}^t [B(c_t) - \lambda C_t] \\ \text{s.t. } C_t &= (1-\rho)C_{t-1} + Mc_t \text{ for all } t \\ C_t &\leq \tilde{C} \text{ for all } t; \;\; c_t \geq 0 \text{ for all } t. \end{aligned}$$

We assume that the problem is globally concave, so that we can rely on a first-order condition approach, and for simplicity we restrict attention to the case where the emissions feasibility constraint $c_t \geq 0$ is not binding.

Here we summarize the main steps of the analysis, relegating the formal proof to the Appendix. There are two cases to consider, depending on whether or not the brink constraint $C_t \leq \tilde{C}$ binds for any t. We show that, if $\tilde{C}$ is higher than a threshold level C^S, the brink constraint is not binding and the brink of collapse $\tilde{C}$ is never reached; we refer to this as Case 1. If $\tilde{C}$ is lower than C^S, on the other hand, the brink constraint is binding and the brink of collapse $\tilde{C}$ is reached at some point in time; we refer to this as Case 2.

[25] Notice that in our setting the planner problem is time-consistent, so we need only write down the planner's objective from the perspective of $t = 0$.

In Case 1, where $\tilde{C} \geq C^S$, we show that the optimal level of emissions $\bar{c}^S$ is constant for all countries and all generations, and defined by $B'(\bar{c}^S) = \frac{M\lambda}{1-\hat{\beta}(1-\rho)}$. This solution has a simple interpretation: a country's marginal benefit from its own emissions should equal the marginal environmental cost of emissions, taking into account the costs imposed on the utility of all M countries and on all future generations (discounted by the planner's discount factor $\hat{\beta}$ and accounting for the natural rate of atmospheric regeneration ρ).

In this case the carbon stock increases in a concave way and converges to the steady state level C^S. In the Appendix we show that C^S is lower than C^N, the steady state level of the BAU global carbon stock. Recalling our Assumption 1 that $\tilde{C} < C^N$, Case 1 therefore obtains if $\tilde{C} \in [C^S, C^N)$. The intuition for this case is the same as for the analogous case in the context of an ICA, namely, if the catastrophe point $\tilde{C}$ is high and sufficiently close to C^N, then only a relatively small reduction in emissions from the BAU level would be required to keep the world from reaching the brink, and the planner will indeed deliver the required reductions.

In Case 1, it is easy to see that the utility of each generation declines through time as C_t rises and the climate warms. It is notable that, while $\hat{\beta}$ impacts the *level* of $\bar{c}^S$, it does not alter the fact that the socially optimal emissions level is constant through time. Evidently, in Case 1 a higher $\hat{\beta}$ induces higher welfare for later generations under the socially optimal emissions choices not by tilting the emissions profile toward later generations, but by reducing the (constant) level of emissions for all generations and thereby shifting utility toward future generations in the form of a lower level of atmospheric carbon and a cooler climate.

In Case 2, where $\tilde{C} < C^S$, the socially optimal carbon stock level grows over time until it reaches the brink level $\tilde{C}$ at some point in time $\tilde{t}^S$. In this case, we show that the socially optimal emissions in a representative country, denoted $\hat{c}_t^S$, decline over time until time $\tilde{t}^S$, at which point they hit the level $\hat{c}^N$ that keeps the carbon stock steady at the brink level $\tilde{C}$, and then remain constant at $\hat{c}^N$.

To gain some intuition for the result that in Case 2 the socially optimal emissions decline over time until they hit their steady state level, consider the special case where $\lambda = \rho = 0$. Then we can think of the problem as allocating a fixed amount of (nonnegative) emissions across all generations, so the optimal emissions maximize $\sum_{t=0}^{\infty} \hat{\beta}^t B(c_t)$ subject to the constraints $\sum_t c_t = \tilde{C}$ and $c_t \geq 0$. Clearly, then, whenever c_t is strictly positive it must equalize the discounted marginal benefit of emissions, $\hat{\beta}^t B'(c_t)$, across generations, so that $\hat{\beta}^t B'(c_t)$ must be constant over time, therefore the undiscounted marginal benefit $B'(c_t)$ must increase, and

hence c_t must decrease over time.

In Case 2, it is easy to see that the level of welfare falls through time for $t < \tilde{t}^S$ – due to the warming climate as in Case 1, but in contrast to Case 1 also due to the decline in consumption implied by the falling emissions – until the brink generation $\tilde{t}^S$ is reached, at which point and contrary to Case 1, both global emissions and the global carbon stock are frozen in place and the decline in welfare is halted thereafter. Also in contrast to Case 1, in Case 2 an increase in the social discount factor $\hat{\beta}$ shifts utility to later generations both by slowing the accumulation of atmospheric carbon and keeping the planet cooler for longer *and* by tilting the emissions profile away from the earliest generations. Finally note that, contrary to the noncooperative and ICA outcomes, when the brink of climate catastrophe is reached under the socially optimal level of emissions, the brink generation does not suffer a precipitous drop in welfare relative to the previous generation.

We summarize the properties of the socially optimal emissions choices with:

Proposition 3. *The socially optimal path of emissions in the common-brink scenario falls into one of two cases. If $\tilde{C}$ is above a threshold level, the brink of catastrophe is never reached and the socially optimal emissions levels are constant through time. Otherwise, if $\tilde{C}$ is below this threshold the socially optimal emissions levels decline through time until the brink of catastrophe is reached, and for the brink generation and all generations thereafter the emissions remain at the replacement rate dictated by the natural rate of atmospheric regeneration. In this second case where the brink is reached, the brink generation does not suffer a precipitous drop in welfare relative to the previous generation.*

3.4. Comparison of ICA and Socially Optimal Outcomes

We now compare the outcomes that are achieved under the ICA with the socially optimal outcomes that would be chosen by the planner. To this end, we begin by noting that we have $C^S < C^{ICA}$ and hence $C^S < C^{ICA} < C^N$. We can thus organize the comparison between the ICA and socially optimal outcomes into three ranges of $\tilde{C}$: high ($\tilde{C} \in [C^{ICA}, C^N)$), intermediate ($\tilde{C} \in [C^S, C^{ICA})$) and low ($\tilde{C} < C^S$).

Consider first the possibility that $\tilde{C}$ falls in the high range $\tilde{C} \in [C^{ICA}, C^N)$. In this case the world will be kept below the brink of climate catastrophe by both the ICA and the planner through the implementation of constant emissions levels $\bar{c}^{ICA}$ and $\bar{c}^S$ respectively that are below

the BAU level $\bar{c}^N$ and that keep the global carbon stock below $\tilde{C}$. However, the planner dictates that the emissions choices $\bar{c}^S$ internalize both international *and* intergenerational effects of those choices, while under the ICA emissions choices $\bar{c}^{ICA}$ only the international climate externalities are internalized; and as a result we have $\bar{c}^S < \bar{c}^{ICA}$, with $\bar{c}^S$ dropping further below $\bar{c}^{ICA}$ as $\hat{\beta}$ increasing and as ρ decreases, and the steady state carbon stock delivered under the ICA is larger than the socially optimal level. The three panels of Figure 1 illustrate the time path of emissions, the global carbon stock, and the utility of a representative country under the ICA and socially optimal emissions as well as in the noncooperative equilibrium for $\tilde{C}$ in this range. The qualitative features of the ICA and socially optimal outcomes are similar, with the difference between the two being that the planner shifts welfare from early generations to later generations relative to the ICA by requiring lower emissions for all generations and thereby reducing the extent to which utility falls through time due to a rising global carbon stock and worsening climate.[26]

Consider next the possibility that $\tilde{C}$ falls in an intermediate range $\tilde{C} \in [C^S, C^{ICA})$. In this case the world would still be kept from the brink of climate catastrophe by the planner, but under the ICA the world will be brought to the brink. This is because with $\tilde{C}$ in this intermediate range, the planner's choice of emissions $\bar{c}^S$ is still low enough to keep the global carbon stock below $\tilde{C}$, but the higher level of emissions $\bar{c}^{ICA}$ implemented during the warming phase of the ICA is no longer low enough to accomplish this. Hence, in this case the inability of the ICA to take into account directly the interests of future generations leads to a qualitative difference across the ICA and socially optimal outcomes. This is reflected in the three panels of Figure 2. As in Figure 1, here the utility of earlier generations is higher and the utility of later generations is lower under the ICA than in the social optimum, but now utility under the ICA falls precipitously for the generation alive when the brink is reached, while under the social optimum the utility of each generation evolves smoothly through time. And while in this case the planner would not let utility for any generation fall to the level of utility experienced in the noncooperative equilibrium by the brink generation, under the ICA the generation alive when the brink is reached and all future generations will experience exactly that level of utility.

[26] We have depicted the level of welfare achieved by early generations in Figure 1 as dropping under the social optimum relative to the noncooperative equilibrium, but this need not be so. If $\hat{\beta}(1-\rho)$ is sufficiently small, the planner will raise the level of welfare achieved by the early generations as well relative to the noncooperative equilibrium, because then the planner is essentially internalizing international but not intergenerational externalities and hence mimics the ICA outcome, which provides (weakly) higher than Nash welfare for every generation.

Finally, consider the possibility that $\tilde{C}$ falls in the low range $\tilde{C} < C^S$. In this case the world will be brought to the brink of climate catastrophe by both the ICA and the planner, but as noted we have $\tilde{t}^N < \tilde{t}^{ICA} < \tilde{t}^S$: the ICA slows down the march to the brink relative to the noncooperative outcome, but this march is still too fast relative to the social optimum. In this case as well there are qualitative differences across the ICA and socially optimal outcomes that arise as a result of the inability of the ICA to take into account directly the interests of future generations. This is reflected in the three panels of Figure 3. Here the ICA emissions remain constant at the level $\bar{c}^{ICA}$ during the warming phase leading up to the brink and then fall precipitously to the level $\hat{c}^N$ for the brink generation, implying an associated precipitous drop in the welfare of the brink generation relative to the previous generation. By contrast, under the socially optimal choices the emissions $\hat{c}_t^S$ during the warming phase decline smoothly over time, and they reach the level $\hat{c}^N$ at the brink without a precipitous drop for the brink generation in either emissions or utility.

Summarizing, we may now state:

Proposition 4. *In the common-brink scenario, the ICA addresses the horizontal (international) externalities that are associated with emissions choices and that create inefficiencies in the noncooperative outcomes, but it cannot address the vertical and diagonal externalities that are associated with the intergenerational aspects of the climate problem. For this reason, the ICA slows down the growth of the carbon stock relative to the noncooperative outcome but not enough relative to the social optimum. More specifically: (i) If $\tilde{C}$ is above a threshold level the ICA prevents the world from reaching the brink of catastrophe, but the steady-state carbon stock is still too large relative to the social optimum. (ii) If $\tilde{C}$ lies in an intermediate range the world reaches the brink of catastrophe under the ICA but not under the social optimum. (iii) If $\tilde{C}$ is below a threshold level the brink is reached both under the ICA and the social optimum, but it is reached faster under the ICA, and under the social optimum the brink generation does not suffer a precipitous drop in welfare.*

It is also natural to wonder how the ICA affects future generations. This is not obvious *a priori*, because for each generation t the ICA is a contract that excludes future generations, and because we are focusing on a scenario without any intergenerational altruism. The answer is that in the common-brink scenario an ICA nevertheless benefits future generations. This is because the act of reducing emissions today under an ICA has two positive effects on future

generations: first, it will leave the next generation with a lower global carbon stock, and hence reduce the environmental losses tomorrow; and second, it will at least to some extent slow down the march to the brink of climate catastrophe, and therefore put off the day of reckoning when emissions and hence consumption levels will need to fall precipitously to save the world.

Finally, returning to Figures 1-3, we may ask which of the three cases depicted in these figures most accurately reflects the true limitations faced by ICAs due to their inability to take into account directly the interests of future generations. According to the common-brink scenario, the answer to this question depends on the severity of the constraint that the catastrophic carbon level $\tilde{C}$ places on attainable steady state welfare. If one takes an agnostic view regarding the relative empirical plausibility of these three scenarios, the message from the model is that the world is more likely to reach the brink of catastrophe under the ICA than it would be under the global social planner; or more specifically, that under the ICA the world reaches the brink of catastrophe for a larger parameter region than under the planner. This is an immediate corollary of Proposition 4, which states that, fixing all other model parameters, the interval of $\tilde{C}$ for which the world reaches the brink is wider under the ICA than under the planner.

But something more can be said if one is willing to rule out a dystopian view of the world in which the planner would find it optimal to allow the world to arrive at the brink of catastrophe and then remain on the brink thereafter, that is, the scenario described by Figure 3. This scenario can be ruled out for any given level of $\tilde{C}$ if the planner's discount factor accounting for the natural rate of atmospheric regeneration, $\hat{\beta}(1-\rho)$, is sufficiently close to one. And while the most optimistic position on the severity of the constraint that $\tilde{C}$ places on attainable steady state welfare would point to Figure 1, recall that this scenario can only apply if the cost associated with moderate degrees of global warming, λ, is above a certain threshold; so if λ is below this threshold, only the middle ground associated with Figure 2 remains.[27] And according to Figure 2, as we have noted, the implications of the inability of ICAs to take into account directly the interests of future generations are especially dire: while a global social planner would keep the world from ever arriving at the brink of climate catastrophe, an ICA will at

[27] More formally, it can be shown that the case in Figure 1 is ruled out if $\lambda < \frac{B'(\frac{\rho\tilde{C}}{M})}{M}$, and the case in Figure 3 is ruled out if $\hat{\beta}(1-\rho) > 1 - \frac{\lambda M}{B'(\frac{\rho\tilde{C}}{M})}$. Assuming $B'(0)$ is finite, and for given values of $\tilde{C}$, M and ρ and fixing a level of $\lambda \in (0, \frac{B'(\frac{\rho\tilde{C}}{M})}{M})$, it then follows that if $\hat{\beta}(1-\rho)$ is close to one – which is possible even under our assumption that $\hat{\beta} < 1$ in the empirically relevant case of ρ close to zero (see note 12) – we are in the case of Figure 2.

best only postpone the arrival at the brink, and when that day arrives, the brink generation and all generations thereafter will suffer a precipitous drop in welfare. Hence we may state:

Corollary 1. *If the cost associated with moderate degrees of global warming, λ, is not too large, and if the planner's discount factor accounting for the natural rate of atmospheric regeneration is close enough to one, the world will reach the brink of catastrophe under the ICA but not under the social optimum.*

Like the fabled boiling frog, Corollary 1 suggests that a slowly rising cost of climate change (small-to-moderate λ) may describe the scenario most likely to cause the world to "remain oblivious" during the warming phase, and thereby arrive at the brink of climate catastrophe under an ICA when a global social planner would not have allowed this to happen. The twist is that, unlike the frog in the fable, the world will not go off the brink under these conditions; but it will be consigned to life on the brink, a fate that the more forward-looking actions of the planner would have avoided.

This is also the case where our model may best capture in a highly stylized way the essence of the plight of the climate activist Greta Thunberg and her generation, and how their plight can be interpreted through the lens of our model as arising from the inability of climate agreements to internalize intergenerational externalities. In an address to world leaders at the United Nation's Climate Action Summit in New York City on September 23 2019, Thunberg stated:

> "You have stolen my dreams and my childhood with your empty words... How dare you! For more than 30 years, the science has been crystal clear... The popular idea of cutting our emissions in half in 10 years only gives us a 50% chance of staying below 1.5 degrees [Celsius], and the risk of setting off irreversible chain reactions beyond human control. Fifty percent may be acceptable to you ... [but it] is simply not acceptable to us — we who have to live with the consequences. To have a 67% chance of staying below a 1.5 degrees global temperature rise – the best odds given by the [Intergovernmental Panel on Climate Change] – the world had 420 gigatons of CO_2 left to emit back on Jan. 1st, 2018. Today that figure is already down to less than 350 gigatons. How dare you pretend that this can be solved with just 'business as usual' and some technical solutions? With today's emissions levels, that remaining CO_2 budget will be entirely gone within less than 8 1/2 years... You are failing us. But the young people are starting to understand your

betrayal. The eyes of all future generations are upon you. And if you choose to fail us, I say: We will never forgive you. We will not let you get away with this. Right here, right now is where we draw the line. The world is waking up. And change is coming, whether you like it or not."

In terms of Figure 2, we might think of the sum total of climate agreements to date as putting the world somewhere between the noncooperative emissions path (if these agreements were completely ineffective) and the ICA path (if the agreements were maximally effective), and we might think of Greta and her generation as corresponding to the brink generation (marked in Figure 2 by $\tilde{t}^N$ in the former case and $\tilde{t}^{ICA}$ in the latter case). The "consequences" to which Greta refers in the quote above are then reflected in Figure 2 by the implication of the threat of a climate catastrophe experienced in her lifetime, a threat caused by the emissions of previous generations that the brink generation must now confront and that might have been avoided if previous generations had adopted socially optimal emissions policies. According to our common-brink scenario, the implication of this threat is that the world will indeed find an 11^{th} hour solution which prevents the threat of climate catastrophe from materializing, much as Greta predicts. But as Figure 2 depicts, avoiding climate catastrophe at the 11^{th} hour comes at the cost of a precipitous drop in utility for the brink generation relative to their parents, and the same low level of utility for all generations thereafter. And this solution comes about not because ICA's will finally find a way to solve the issues that have bedeviled cooperation over climate issues for decades, but because the consequences of *not* finding a solution are so dire for the brink generation that the (intergenerational) externalities that limited the attempts of previous generations to address the problem are no longer an impediment to its solution, and at that point an ICA is not even needed (or is needed only to help avoid a coordination failure).

4. The Heterogeneous-Brink Scenario

In the previous section we considered a special case of the modeling framework presented in section 2 where the level of the carbon stock that would be catastrophic, $\tilde{C}$, is assumed to be the same for all countries. And under this assumption, we argued that even in a noncooperative equilibrium catastrophe will be averted, because countries will find a way to do whatever it takes to prevent mutual collapse. But what if the global carbon stock at which a catastrophe would be triggered differs from one country to the next? And what if not all (or even any)

countries have the capacity to avoid collapse on their own? For example, it is often observed that small island nations such as the Maldives are especially vulnerable to the effects of climate change and may soon face an existential threat posed by rising sea levels.[28] If some countries face existential threats from climate change before others, new questions arise. Under what conditions will some (or even all) of the countries collapse on the noncooperative equilibrium path? Can there be domino effects, where the collapse of one country hastens the collapse of the next? If some or all countries would collapse in the noncooperative equilibrium, can ICAs help to avoid collapse? And what is the outcome that a global social planner would implement in this case?

To answer these questions, we now move beyond the common-brink benchmark scenario considered in section 3 and consider the more general scenario described in section 2, where countries are allowed to reach a catastrophe at different levels of the global carbon stock. Recall that $\tilde{C}_i$ denotes the level of the carbon stock beyond which country i collapses. We order countries according to increasing $\tilde{C}_i$, so that country 1 is the country with the lowest $\tilde{C}_i$ and therefore the country "least resilient" to climate change, while country M has the highest $\tilde{C}_i$ and is hence the most resilient country. We also assume for simplicity that this ordering is strict, i.e. no two countries have the same value of $\tilde{C}_i$.[29]

Recall from section 2 that there are two costs associated with a country's collapse. The first is a one-time per-capita utility cost L suffered by the citizens of the collapsing country. We assume collapse occurs at the end of a period, after consumption has occurred, and we will think of L as high but finite as long as there are other surviving countries to which the citizens of the collapsing country can immigrate; we will only think of L as infinite for the citizens of the *last* surviving country who, as in the common-brink scenario, facing collapse would have nowhere to go.[30] The second cost of a country's collapse is borne by the remaining

[28] And it is not just through rising sea levels that global-warming induced climate changes can have differential effects on countries. For example, Lenton et al. (2008) describe a tipping point exhibited by global-warming-induced changes in the amplitude of the El Nino - Southern Oscillation that would trigger drought conditions in Southeast Asia. See also Jones and King (2021), who develop a methodology for predicting a shortlist of countries that are most likely to be left standing when other countries have succumbed to climate catastrophe.

[29] While we focus formally on heterogeneity *across* countries, it is worth noting that many of the same issues that we consider below will arise *within* countries, if there is heterogeneity in climate collapse points across different regions due to distinct geographical and/or socioeconomic features. Such regional heterogeneity raises issues for federal versus regional government emissions policy choices that are analogous to the issues we identify below for global planner/ICA versus noncooperative national emissions policy choices (see for example Lustgarten, 2020). We leave an exploration of these parallel themes to future work.

[30] In assuming that the cost borne by each citizen of a collapsing country is a constant L as long as there remain other surviving countries to which the citizens of the collapsing country can immigrate, we are abstracting from

countries. The collapsing country's citizens become climate refugees (spreading equally across the remaining countries of the world), and each climate refugee imposes a one-time utility cost r on the country to which it immigrates.

Letting H_t index the most vulnerable country that has survived to time t, the number of surviving countries at t is $M - H_t + 1$, and the population of a surviving country at t is $\frac{M}{M-H_t+1}$. If country H_t collapses, since the total population of the remaining countries is $M - \frac{M}{M-H_t+1} = \frac{M(M-H_t)}{M-H_t+1}$, it follows that the one-time per-capita utility cost incurred by citizens of the remaining countries as a result of country H_t's collapse is

$$R_{H_t} \equiv \frac{r}{M - H_t}.$$

As with L, we assume that the refugee cost R_{H_t} is incurred at the end of the period of country H_t's collapse.[31] Notice also that the refugee externality R_{H_t} is increasing in H_t. This is because in our model countries that collapse later (higher H_t) release a greater number of climate refugees ($\frac{M}{M-H_t+1}$) on a smaller rest-of-world population ($M - \frac{M}{M-H_t+1}$). But as will become clear below, our results would be unchanged if the refugee externality were independent of H_t.

Recall that in the common-brink scenario of section 3, the population of each country remained constant over time and we normalized this population to one, so country-level and per-capita-level variables were one and the same. But with climate refugees altering the population of surviving countries when more vulnerable countries collapse, country-level and per-capita-level variables will diverge, so we now specify $u_{i,t}$ at the per-capita level, and in particular as the utility of a person living in country i in period t. We will continue to interpret $c_{i,t}$ as the per-capita emissions of country i in period t, and $B(c_{i,t})$ as the per-capita benefit from consumption that comes from emitting at the per-capita level $c_{i,t}$.

To preserve tractability, we assume that both L and R_{H_t} enter utility in an additively separable way (so that they do not impact directly the emissions choices of countries and instead act as simple shifters of utility). With this assumption, the utility of a citizen living in country i at time t is given by:

$$u_{i,t} = B(c_{i,t}) - \lambda C_t - L \cdot E_{i,t} - R_{H_t} \cdot I_{i,t}, \tag{4.1}$$

the possibility that this cost might rise as more countries collapse and there are fewer surviving countries in which climate refugees can resettle. Incorporating this possibility into the model would be straightforward, but it would not alter the results we emphasize below, so we choose to opt for simplicity and leave it out.

[31] Since time-periods correspond to generations in our model, it seems reasonable to assume that L and R_{H_t} are incurred as one-time costs, but at the end of this section we also discuss briefly the case of permanent per-period costs.

where $E_{i,t}$ is an indicator function that equals one if country i collapses at time t, so that its population has to emigrate, and $I_{i,t}$ is an indicator function that equals one if some country other than country i collapses at time t, so that country i receives immigrants from the collapsing country.[32] And of course, the utility function in (4.1) is defined only for the countries that have survived up to time t.

Even absent outright collapse, the link between climate change and refugees can already be seen for example in the recent surge of migrants at the southern U.S. border, as Russonello (2021) writes in the *New York Times*:

> "The main motivators of emigration from Mexico, Central America and points south are tied to climate change, violent crime and corruption – all issues that the Biden administration knows it must confront if it stands any chance of stemming the inflow of people at the border. (...) The most immediate cause of the immigration surge may be the series of deadly hurricanes that swept through Central America last year, part of a greater trend fueled by climate change. They destroyed crops and homes, especially in Honduras, leaving an estimated nine million people displaced. They've had six years of ongoing droughts in these areas, they have no food, no means for employment or livelihood, and they're eating the seeds which they would normally save for planting."

Our reduced-form modeling of the costs of a country's collapse is meant to capture the costs that are incurred when a climate catastrophe makes a country uninhabitable and its citizens must seek residency elsewhere.

Finally, the carbon stock in our heterogeneous-brink scenario evolves according to

$$C_t = (1-\rho)C_{t-1} + \sum_{i=H_t}^{M} \frac{M}{M-H_t+1} \cdot c_{i,t} \text{ with } C_{-1} = 0. \tag{4.2}$$

Notice that according to (4.2) we now have $\frac{\partial C_t}{\partial c_{i,t}} = \frac{M}{M-H_t+1}$: the impact of a surviving country's per-capita emissions on the carbon stock C_t grows as the number of surviving countries shrinks, because country population grows due to the absorption of climate refugees.

[32] In writing (4.1) we have implicitly assumed that no more than one country could collapse in a given period t, but it is straightforward to accommodate the possibility of multiple-country collapses in a given period at the expense of additional notation.

4.1. Noncooperative Equilibrium

We first characterize the noncooperative emissions choices. While in principle we now have two state variables, C_{t-1} and H_t, it is clear that the identity of the most vulnerable country that has survived to time t, H_t, is pinned down by the previous period's carbon stock C_{t-1}; so we can continue to regard C_{t-1} as the only state variable. And as in the previous section, with $\beta = 0$ countries are effectively myopic and so for each level of C_{t-1} we effectively have a one-shot game, and hence we can continue to focus on (subgame perfect) Nash equilibria (or simply the "equilibria").

Given $\beta = 0$, it is easy to see that along the noncooperative path the world may pass through three possible phases: a warming phase, where warming takes place but no catastrophes occur; a catastrophe phase, where warming continues and a sequence of countries collapse; and a third phase where warming and catastrophes are brought to a halt. The first and third phases are familiar from the analysis of the common-brink scenario; the possibility of a middle phase in which some countries collapse along the noncooperative path is novel to the current setting where catastrophe points differ across countries. Notice, too, that with $\frac{\partial C_t}{\partial c_{i,t}} = \frac{M}{M-H_t+1}$, the BAU per-capita emissions of a surviving country now depend on H_t, so we write them as $\bar{c}^N_{H_t}$. These are implicitly defined by

$$B'(\bar{c}^N_{H_t}) = \frac{M}{M - H_t + 1} \cdot \lambda. \tag{4.3}$$

From (4.3), the BAU per-capita emissions is the same across all surviving countries (so we can omit the country subscript i) but falls as the number of surviving countries shrinks (H_t rises) and the population of each surviving country rises. This simply reflects the fact that with each country collapse there are fewer remaining countries in the world; and the countries that do remain internalize a greater proportion of the global cost of their BAU emissions choices. Since the world population remains constant at M, the world BAU emissions level $M\bar{c}^N_{H_t}$ also falls with each country collapse.

To avoid uninteresting taxonomies we assume that, if a country is at its brink and there are other surviving countries, the former is not able to "save itself" by cutting its own emissions to zero if the latter choose their BAU emissions. In essence we are assuming that a country is not able to unilaterally stop the growth of the global carbon stock unless it is the lone surviving country, a feature that seems empirically plausible. Formally, since the BAU emissions $\bar{c}^N_{H_t}$

decline as H_t increases, a sufficient condition for this assumption to hold is

$$\frac{M}{2}\bar{c}^N_{M-1} > \rho\tilde{C}_{M-1}, \qquad \text{(Assumption 2)}$$

a restriction that we will maintain throughout.[33]

To develop some intuition for how the noncooperative path is determined in this setting, it is useful first to focus on the case where $r = 0$ so that there are no refugee externalities. In this case the equilibrium path of the noncooperative game is simple, and it provides a sharp counterpoint to the equilibrium of the noncooperative game in the common-brink scenario of the previous section.

After an initial warming phase during which there are no catastrophes and each country selects the BAU emissions level $\bar{c}^N_1$, the world enters a catastrophe phase when country 1 arrives at the brink of collapse. This occurs in finite time if $M\bar{c}^N_1 > \rho\tilde{C}_1$, which is implied by Assumption 2. Furthermore, Assumption 2 implies that country 1 is not able to offset the BAU emissions of the rest of the world and save itself – and with $r = 0$, the remaining countries have no reason to help country 1 survive – and hence country 1 will collapse. And by a similar logic, all countries except for the most resilient one ($i = M$) will collapse on the equilibrium path. The most resilient country is guaranteed to survive on the equilibrium path, because by setting $c = \frac{\rho\tilde{C}_M}{M} \geq 0$ it can freeze the global carbon stock at the brink level $\tilde{C}_M$.[34]

Hence, with heterogeneous brinks and $r = 0$ all countries except the least vulnerable one will collapse along the equilibrium path, provided only that, as we have assumed, no individual country is able to fully offset the other countries' BAU emissions. This scenario provides an illuminating contrast to the common-brink case, where once the world reached the brink of catastrophe countries did whatever was necessary to avoid global collapse. Relative to that setting, the difference here is that each country has its *own* brink generation, who faces the existential climate crisis *alone* and up against the other countries, who do not internalize the impact of their emissions on the fate of the brink country. Notice also that even slight differences in collapse points across countries can lead to collapses along the equilibrium path: unless each

[33] To understand Assumption 2 note that, when $H_t = M - 1$, there are only two surviving countries with population $M/2$ in each, so if country M chooses its BAU per-capita emissions $\bar{c}^N_{M-1}$ and country $M-1$ chooses zero emissions, total world emissions are $(M/2)\bar{c}^N_{M-1}$, and if this level exceeds $\rho\tilde{C}_{M-1}$ then country $M-1$ will collapse once its brink is reached. And if country $M-1$ is not able to "save itself," neither can the more vulnerable countries ($k < M-1$).

[34] Recall that we have assumed that L is large enough (infinite) when country M is the last surviving country to induce it to do whatever is necessary to avoid its own collapse.

country arrives at the brink at the same time, the "we are all in this together" forces that enabled the world to avoid collapse in the noncooperative equilibrium of our common-brink scenario will be disrupted. As we next demonstrate, allowing for climate refugee externalities will bring back an element of these forces, albeit only partially.

To proceed, we now allow for climate refugee externalities ($r > 0$). The game can be solved in two steps. First, for each level of C_{t-1} we characterize the equilibrium period-t emissions choices $c_i^N(C_{t-1})$ in each surviving country. And second, we derive the implied equilibrium path for C_{t-1} and hence for the set of countries that survive to each t. As with our analysis of the common-brink scenario, in what follows we ignore integer constraints.

Consider first what happens in a period t if in that period no country is on the brink of catastrophe ($C_{t-1} \neq \tilde{C}_k$ for all k). In this case, clearly each country chooses its BAU emissions $\bar{c}_{H_t}^N$ defined by (4.3).

Next consider what happens in a period t if at the beginning of that period some country k is on the brink of catastrophe ($C_{t-1} = \tilde{C}_k$). Countries $j \geq k$ have survived up to this point, each with a population of $\frac{M}{M-k+1}$. Clearly, then, country k will survive if and only if $\frac{M}{M-k+1} \sum_{j=k}^{M} c_j \leq \rho \tilde{C}_k$. We now argue that there are two possible types of equilibrium.

The first possibility is that all countries choose the BAU emissions level $\bar{c}_k^N$ and country k does not survive. This is always an equilibrium, because given Assumption 2 no country (including country k itself) could unilaterally save country k if the other countries choose $\bar{c}_k^N$. Also note that, if the other countries choose $\bar{c}_k^N$, country k's best response is to also choose $\bar{c}_k^N$, because it gets to enjoy the benefit of its current-period emissions before it collapses, and therefore, given that collapse is inevitable, it can do no better than to choose its BAU emissions.

The second possibility is that country k survives, and the countries' emissions levels satisfy $\frac{M}{M-k+1} \sum_{j=k}^{M} c_j = \rho \tilde{C}_k$. Intuitively, this type of equilibrium can exist only if the refugee externality is large enough, so that countries $j > k$ have an incentive to "top off" the mitigation efforts of country k, ensure country k's survival, and avoid a climate refugee crisis. It is easy to show that equilibria with survival of country k exist if and only if it is an equilibrium for country k to do everything feasible to save itself by setting $c_k = 0$ and for the remaining countries ($j > k$) to top off this effort in a symmetric way.[35] Note that, given $c_k = 0$, the maximum symmetric

[35] To see this, note that (i) raising c_k above zero would require reducing emissions for some country $j > k$, and this would increase the latter country's temptation to defect; and (ii) assigning asymmetric emissions to countries $j > k$ would make it harder to sustain such an equilibrium, because it would increase the temptation to defect for the countries that have a bigger burden.

emissions level for each country $j > k$ that ensures survival of country k is $\frac{\rho\tilde{C}_k}{M-k}$, and since the population of each country at this stage is $\frac{M}{M-k+1}$, the corresponding per capita emissions level is $\frac{M-k+1}{M} \cdot \frac{\rho\tilde{C}_k}{M-k} \equiv c_k^{save} < \bar{c}_k^N$. We will refer to this as the "self-help" equilibrium. This is arguably the most intuitive equilibrium among those in which country k survives, because it minimizes the less vulnerable countries' gain from deviating. In what follows, if there are equilibria where country k survives, we will restrict our focus to the self-help equilibrium, but our main results do not depend on this equilibrium selection assumption.[36]

We now write down the conditions under which the self-help equilibrium described above arises. The no-defect condition for a country $j > k$ can be written as follows:

$$G_k \equiv \left[B(\bar{c}_k^N) - \lambda\left(\tilde{C}_k + \frac{M}{M-k+1} \cdot \left(\bar{c}_k^N - c_k^{save}\right)\right)\right] - \left[B(c_k^{save}) - \lambda\tilde{C}_k\right] \leq R_k. \tag{4.4}$$

The left-hand side of (4.4) is the gross per-capita gain to a country $j > k$ were it to defect from the self-help equilibrium and cause country k to collapse. The term in the first square brackets is the per-capita payoff to country $j > k$ from deviating to the BAU emissions level $\bar{c}_k^N$ and causing country k to collapse; the term in the second square brackets is the per-capita payoff to country $j > k$ under the self-help equilibrium emissions levels in which country k does not collapse. Note that, when defecting, country j's emissions increase by the amount $\frac{M}{M-k+1}\left(\bar{c}_k^N - c_k^{save}\right)$, and this causes the carbon stock to exceed $\tilde{C}_k$ by the same amount. The right-hand side is the per-capita cost that country $j > k$ incurs as a result of k's collapse. Thus, as intuition suggests, an equilibrium with survival of country k exists only if the refugee externality R_k that country k's collapse would impose on the remaining countries is large enough.

Turning to the no-defect condition for country k, this condition can be written as:

$$G_k^0 \equiv \left[B(\bar{c}_k^N) - \lambda\left(\tilde{C}_k + \frac{M}{M-k+1} \cdot \bar{c}_k^N\right)\right] - \left[B\left(0\right) - \lambda\tilde{C}_k\right] \leq L. \tag{4.5}$$

The left-hand side of (4.5) is country k's per-capita gain in defecting from the self-help equilibrium, in terms of the additional welfare it enjoys for the period up until the moment of its collapse. The term in the first square brackets is the per-capita welfare that country k would

[36] It is worth noting that the self-help equilibrium does not maximize the joint payoff of the surviving countries, since it does not equalize their marginal benefit of emissions ($B'(\cdot)$); but also note that the equilibria with survival of country k (if they exist) are not Pareto-rankable, since international transfers are not used in a noncooperative equilibrium. This suggests that, if countries indeed focus on the self-help equilibrium, one of the potential roles of an ICA will be to allow countries to move to the efficient allocation of emissions through the use of transfers, analogous to the potential ICA role that we describe under the common-brink scenario in note 23. We will come back to this point in the next subsection.

enjoy were it to deviate to the BAU emissions level $\bar{c}_k^N$, and the term in the second square brackets is its per-capita welfare under the self-help equilibrium emissions levels. The term on the right-hand side is the per-capita cost that the collapse of country k, precipitated by its own defection from the self-help equilibrium, would impose on its citizens at the end of the period.

An equilibrium with survival of country k exists if and only if both (4.4) and (4.5) are satisfied. Notice that, if this is the case, the equilibrium where country k collapses (described as the first possibility above) is Pareto-dominated.[37] It can also be shown that there are no other possible equilibria beyond the two types of equilibria we have just described. Thus, maintaining our emphasis on Pareto-undominated equilibria as we did in the analysis of noncooperative equilibria in the common-brink scenario (we will again comment in a later section on the role of ICAs in addressing possible coordination failures), we can conclude that if country k is at the brink of catastrophe in period t ($C_{t-1} = \tilde{C}_k$), this country survives in equilibrium if and only if (4.4) and (4.5) are satisfied.

Finally note that, if $C_{t-1} = \tilde{C}_M$, so that the only surviving country $j = M$ is at the brink of catastrophe in period t, this country will restrain its emissions just enough to avoid collapse. Since at this stage the entire world population M is located in this country, its per capita emissions are given by $c_M^N(\tilde{C}_M) = \frac{\rho \tilde{C}_M}{M}$. If the world reaches this stage, the carbon stock stops growing and will stay at the level $\tilde{C}_M$ forever thereafter.

Our next observation is that the gains from defection from the self-help equilibrium, G_k as defined by the left-hand side of (4.4) and G_k^0 as defined by the left-hand side of (4.5), decrease with the number of countries that have collapsed in the past, and hence with k. Consider first G_k. Intuitively, as more countries collapse, the gain from defection G_k becomes smaller, for three reasons. First, there are fewer surviving countries and hence more people in any country $j > k$ that considers defection, implying that a defection has greater (negative) consequences for the global climate. Second, there are fewer surviving countries and hence more people in the

[37] To see this, consider first country k. Since (4.5) is satisfied, country k prefers to emit zero and survive rather than deviating from the self-help equilibrium, emitting $\bar{c}_k^N$, and collapsing at the end of the period. But k's payoff would be smaller still if all countries emitted at the level $\bar{c}_k^N$, as would be true in the equilibrium where country k collapses, because the carbon stock would be larger under the latter scenario. So country k is better off in the self-help equilibrium than in the equilibrium where country k collapses. What about a country $j > k$? The argument is similar. If the self-help equilibrium exists, then we must have (4.4), and hence country j prefers the self-help equilibrium to deviating from c_k^{save}, emitting $\bar{c}_k^N$, and having country k collapse at the end of the period. But j's payoff would be smaller still if all countries emitted at the level $\bar{c}_k^N$ as would be true in the equilibrium where country k collapses, because the carbon stock would be larger under the latter scenario. So country $j > k$ is better off in the self-help equilibrium than in the equilibrium where country k collapses. The claim then follows.

"brink country" (k) emitting zero carbon, and this increases c_k^{save}; and third, $\tilde{C}_k$ increases, and this also increases c_k^{save}. The fact that c_k^{save} increases through these last two channels means that countries $j > k$ can afford to emit more while still saving country k, thus the sacrifice needed to save country k becomes smaller. A similar argument applies to G_k^0, except that only two of the three effects described above are at work, since c_k^{save} is replaced by 0.[38]

Since G_k and G_k^0 are decreasing in k and R_k is increasing in k, we can say that, conditional on a given country k being at the brink, this country is more likely to survive if more countries have collapsed in the past. The following lemma summarizes the key features of the equilibrium outcome conditional on country k being on the brink:

Lemma 1. *Conditional on country k being on the brink in period t ($C_{t-1} = \tilde{C}_k$): (i) If (4.4) and (4.5) are satisfied, country k survives with the help of emissions reductions below BAU levels by other countries; (ii) If (4.4) or (4.5) is violated, country k collapses and the surviving countries continue to choose their BAU emissions. (iii) Country k is more likely to survive if more countries have collapsed in the past.*

The above analysis raises an interesting question: Does the heterogeneous-brink scenario exhibit a "domino effect" when countries collapse on the equilibrium path? At one level the answer is yes, for the simple reason that if a country at the brink is able to avert its own collapse (with the help of other countries), no more dominos will fall. In other words, a given country i can reach the brink only if all the countries that are more vulnerable than country i (that is countries $1, 2, ..., i-1$) have already collapsed. In this sense the heterogeneous-brink scenario exhibits a domino effect. However, as Lemma 1 and the preceding discussion highlights, *conditional on a country reaching the brink* the likelihood of collapse is lower if more countries have collapsed in the past, so in this sense there is also an "anti-domino effect" in the

[38] Formally, rewrite G_k as $G_k = \left[B(\bar{c}_k^N) - \lambda \frac{M}{M-k+1}\bar{c}_k^N\right] - \left[B(c_k^{save}) - \lambda \frac{M}{M-k+1}c_k^{save}\right]$. Note that the term in the first square brackets is maximized by $\bar{c}_k^N$, so we can apply the envelope theorem and write:

$$\frac{dG_k}{dk} = -\lambda \frac{M}{(M-k+1)^2}\left(\bar{c}_k^N - c_k^{save}\right) - \frac{d}{dc_k^{save}}\left(B(c_k^{save}) - \lambda \frac{M}{M-k+1}c_k^{save}\right) \cdot \frac{dc_k^{save}}{dk}. \tag{4.6}$$

The first term on the right-hand side of (4.6) is negative, since $\bar{c}_k^N > c_k^{save}$. Turning to the second term and recalling that $c_k^{save} = \frac{M-k+1}{M} \cdot \frac{\rho \tilde{C}_k}{M-k}$, it is easy to check that $\frac{dc_k^{save}}{dk} > 0$. Next note that $\frac{d}{dc_k^{save}}(\cdot) > 0$ because c_k^{save} is lower than $\bar{c}_k^N$, the emissions level that maximizes $B(c) - \lambda \frac{M}{M-k+1}c$. This ensures that the second term on the right-hand side of (4.6) is negative as well. We can conclude that the derivative in (4.6) is negative, and hence $\frac{dG_k}{dk} < 0$ as claimed. A similar argument can be applied to show that also G_k^0 is decreasing in k.

heterogeneous-brink scenario.[39]

Having characterized the equilibrium emissions conditional on the global carbon stock C_{t-1}, it is straightforward to back out the implied equilibrium path for C_{t-1} and hence for the set of countries that survive to each t. In the initial phase, all countries are present and the growth of C_{t-1} is dictated by the BAU emissions level $\bar{c}_1^N$. Once C_{t-1} reaches the level $\tilde{C}_1$ and country 1 is put in danger in period t, if the refugee externality that the collapse of country 1 would exert on a representative citizen of the rest of the world is not severe enough ($R_1 < G_1$), then country 1 collapses at the end of period t and the rest of the world carries on with their BAU emissions level $\bar{c}_2^N$. In a similar fashion, the growth of the carbon stock will cause the sequential collapse of further countries.

Note that our model allows for the internal cost of collapse L to be low enough that country 1 itself prefers to collapse rather than cutting emissions to zero, which would *a fortiori* imply that country 1 collapses in equilibrium, but empirically it seems plausible that L is high enough that the no-defect condition for the country at the brink (4.5) is not binding. For this reason we emphasize this case in our discussion here, but all of our results are valid more generally.

The condition $R_1 < G_1$, under which a subset of countries collapses on the noncooperative equilibrium path, is arguably quite weak. In reality, the negative externality felt by other countries if a country like the Maldives suffers an early collapse will be limited, due both to the relatively small population of climate refugees that would be released by these countries and to the fact that the associated refugee externality triggered by this early collapse would be shared across many countries.

When does the string of catastrophes end? Recall that the refugee externality R_k increases as more and more countries collapse, while the gain from defecting G_k decreases, thus the process will stop when either (i) R_k rises above G_k, at which point the surviving countries become proactive and help the country at the brink avoid collapse, or (ii) country M becomes the lone surviving country, in which case this country will take care of itself and stop the growth of the carbon stock. In either case, we can state the following:

Proposition 5. *Suppose the catastrophe point $\tilde{C}_i$ differs across countries: (i) If the refugee*

[39] There are also forces outside the heterogeneous-brink scenario that would push in the direction of an anti-domino effect, namely: (a) if part of the collapsing country's population perishes, the total world population will fall, and this will push in the direction of lower aggregate BAU emissions; and (b) to the extent that other resources, such as land and capital, are lost when a country collapses, this will push further in the same direction.

externality imposed by the collapse of the most vulnerable country on the remaining countries is not severe enough, then a non-empty subset of countries will collapse on the noncooperative equilibrium path. This is true even if the differences between catastrophe points ($\tilde{C}_i$) across countries are small. (ii) A given country i can reach the brink only if the countries that are more vulnerable (countries $1, 2, ..., i-1$) all have collapsed (a basic "domino effect"). But the likelihood of country i surviving in period t conditional on having reached the brink in period $t-1$ ($C_{t-1} = \tilde{C}_i$) is higher if more countries have collapsed before it (an "anti-domino effect").

Note the contrast with the common-brink scenario, where catastrophe never happens on the equilibrium noncooperative path. When asymmetries in the collapse points are introduced, the result changes dramatically, and equilibrium catastrophes become likely; moreover, the conditions under which a given country collapses on the equilibrium path are not affected by the distance between the catastrophe point of this country and those of other countries, as long as the catastrophe points are different, so the asymmetries need not be large.

It is also notable that, when collapse points are heterogeneous, it is entirely possible that some countries could continue to enjoy a reasonable level of utility once the carbon stock has stabilized while others have suffered climate collapse. Hence, the heterogeneous-brink scenario brings into high relief the possibly uneven impacts of climate change across those countries who, due to attributes of geography and/or socioeconomic position, are more or less fortunate.

4.2. International Climate Agreements

We next revisit the potential role for ICAs, but now in the setting of the heterogeneous-brink scenario where the catastrophe point differs across countries. In the case of symmetric countries analyzed in the common-brink scenario of the previous section there was no role for international transfers, so we abstracted from them. But in the present setting where collapse points are heterogeneous across countries, international transfers become a relevant consideration. Moreover, such transfers play a prominent role in real world discussions of approaches to address climate change through international agreements (see, for example, Mattoo and Subramanian, 2013), and allowing them therefore seems important. So in the context of ICAs (and later, the social optimum) we now introduce international transfers explicitly into the model. In our formal analysis we will assume that there is no limit on the potential size of these transfers, so that we can continue to focus on the inability of the ICA to take the interests of future generations directly into account as the source of potential shortcomings of ICA outcomes relative to the

social optimum. We will also comment, however, on how our results would be affected if the size of international transfers were limited by resource constraints.

Formally, we model international transfers as lump-sum transfers of an outside good that enters additively into utility. Each country is endowed with the same amount of this outside good each period. To keep things simple, we assume that the endowment is large enough that it never imposes a binding constraint on transfers for any country. We denote by $Z_{i,t}$ the (positive or negative) per-capita transfer made by country i at time t. The utility of a citizen living in a (surviving) country i at time t can thus be written, building on (4.1), as $\hat{u}_{i,t} \equiv u_{it} - Z_{it}$ (where we omit the fixed endowment of the outside good from the utility function to simplify notation). Note that the absence of intergenerational transfers implies $\sum_{i=1}^{M} Z_{i,t} = 0$ for all t.

We are now ready to analyze the equilibrium path of carbon emissions and the carbon stock under an ICA, and derive the implications for the collapse and survival of countries. For a given generation t, the ICA specifies emissions for each country that has survived to date in order to maximize world welfare and then uses international transfers to divide up the surplus across countries according to their bargaining powers, with the noncooperative equilibrium serving as the threat point for ICA negotiations. We keep the determination of international transfers in the background, and focus below on determining the emissions levels that maximize world welfare. As all countries are symmetric except for the threshold $\tilde{C}_i$, efficiency again dictates that the ICA choose the same per-capita emissions in period t for each country that has survived to that point, so as before we omit the country subscript i on ICA emissions.

Under the assumption that $\beta = 0$ and recalling that the population of country H_t, the most vulnerable country that has survived to time t, is $\frac{M}{M-H_t+1}$ while the population of the remaining countries is $M - \frac{M}{M-H_t+1} = \frac{M(M-H_t)}{M-H_t+1}$ and using $R_{H_t} = \frac{r}{M-H_t}$, we can then write the average per-capita world welfare at time t as

$$U_t = B(c_t) - \lambda C_t - \left(\frac{L+r}{M-H_t+1}\right) \cdot I_t \tag{4.7}$$

where I_t is an indicator function that equals one if country H_t collapses at time t and zero otherwise (and recalling our assumption that parameters are such that at most one country collapses at a given time t). The ICA for generation t will choose c_t to maximize U_t.

Consider first the initial warming phase, in which the carbon stock is below the catastrophe point for country 1 ($C_{t-1} < \tilde{C}_1$). In this phase the ICA selects the symmetric level of emissions that maximizes the common per-period payoff, which is given by $\bar{c}^{ICA} \equiv B'^{-1}(M\lambda)$, just as in

the warming phase of the ICA in the common-brink scenario analyzed in the previous section: as before, the ICA internalizes the international climate externalities that travel through λ, and hence lowers emissions below the BAU level $\bar{c}_1^N = B'^{-1}(\lambda)$.

Can the warming phase go on forever under the ICA? Recall that under our Assumption 2, country 1 will reach the brink of catastrophe under the noncooperative scenario (and may or may not collapse), so the warming phase must end in the absence of an ICA. And from our analysis of the common-brink scenario, we know that $C^{ICA} \equiv \frac{M}{\rho} B'^{-1}(M\lambda)$ is the level to which the carbon stock would eventually converge given the emissions level $\bar{c}^{ICA}$. It therefore follows that if $\tilde{C}_1 \geq C^{ICA}$, the carbon stock never reaches $\tilde{C}_1$ under the ICA and so country 1 is never brought to the brink of catastrophe and ICA emissions remain at the level $\bar{c}^{ICA}$ forever. Thus, in the case where $\tilde{C}_1 \geq C^{ICA}$ we may conclude that in contrast to the noncooperative outcome the ICA prevents any country from ever reaching the brink and the warming phase can go on forever. On the other hand, if $\tilde{C}_1 < C^{ICA}$, country 1 is brought to the brink under the ICA, just as it would be in the ICA's absence, and we need to consider what happens next.

Suppose, then, that $\tilde{C}_1 < C^{ICA}$, and country 1 has reached the brink of catastrophe under the ICA. To avoid a taxonomy of uninteresting cases, we focus on the case in which, absent an ICA, a non-empty subset of the most vulnerable countries would collapse. That is, if country $\bar{k}^N$ denotes the marginal surviving country in the noncooperative equilibrium, we are focusing on the case $\bar{k}^N > 1$, which recall is guaranteed under the rather weak condition $R_1 < G_1$.

In this case, when country 1 reaches the brink of catastrophe it would collapse in the absence of an ICA, and the generation alive in the world at this moment now faces a very different international cooperation problem than the problem faced by previous generations. In particular, the world is now confronted with a stark choice: it can cooperate to save country 1 from collapse, or it can let country 1 collapse at the end of period t and carry on without it.

The availability of international lump-sum transfers ensures that the ICA will make this choice so as to maximize the average per-capita world welfare from the point of view of generation t as defined in (4.7). This implies that country 1 will be saved under the ICA if and only if the global loss from the collapse of country 1 exceeds the (minimum) cost to the world of cutting emissions by the sufficient amount to stop the growth of the carbon stock; or put differently, country 1 will be saved if and only if it would be willing to compensate the rest of the world for contributing to stop the growth of the world carbon stock at the level $\tilde{C}_1$.[40]

[40] Given the assumption that countries differ only in the catastrophe threshold $\tilde{C}_i$, in our stylized model

The global average per-capita loss from the collapse of country 1 is comprised of country 1's own loss L (the per-capita utility cost borne by the citizens of country 1 times its normalized initial population of one) and the refugee externality r that its collapse (and, in light of its normalized initial population, its release of one refugee) would impose on others, averaged over the world population M. And recalling that under our assumptions collapse occurs at the end of a period after all consumption for the period has occurred, it follows that if country 1 is allowed to collapse the ICA will nevertheless implement the country-level emissions $\bar{c}^{ICA}$ for that period. On the other hand, the efficient way to save country 1 is for all countries (including country 1) to reduce per capita emissions to the level $\rho\tilde{C}_1/M$, since efficiency requires the marginal benefit from emissions to be equalized across countries.

More generally and with the above logic in mind, if a given country k is at the brink of catastrophe under the ICA, it will be saved if and only if

$$\Gamma_k \equiv \left[B(\bar{c}^{ICA}) - \lambda\left((1-\rho)\tilde{C}_k + M\bar{c}^{ICA}\right)\right] - \left[B\left(\frac{\rho\tilde{C}_k}{M}\right) - \lambda\tilde{C}_k\right] \leq \frac{L+r}{M-k+1} \equiv \psi_k \quad (4.8)$$

The left-hand side of (4.8), Γ_k, is the difference in gross per-capita welfare between (i) having all countries continue to emit at the level $\bar{c}^{ICA}$ and letting country k collapse at the end of the period, and (ii) having all countries emit at the per capita level $\rho\tilde{C}_k/M$ and saving country k from collapse. The right hand side of (4.8), ψ_k, is the global average per-capita refugee cost imposed on the world by the collapse of country k.

The marginal surviving country under the ICA, which we denote $\bar{k}^{ICA}$, is the lowest k such that condition (4.8) is satisfied. Notice that ψ_k is increasing in k, for essentially the same reasons that the refugee externality R_k is increasing: each climate refugee bears a one-time utility cost L and imposes a one-time utility cost r on the country to which it migrates, and countries that collapse later (higher k) release a greater number of climate refugees ($\frac{M}{M-k+1}$) on a smaller rest-of-world population ($M - \frac{M}{M-k+1}$). Thus the global cost imposed by the collapse

the only possible role for international transfers is for more vulnerable countries to compensate more resilient countries, as highlighted just above. But in a richer model where countries differ in other dimensions as well, international transfers could run in different directions. This point can be easily understood even in a setting without the possibility of catastrophes. For example, suppose some countries are poorer than others, and the poorer countries have a higher marginal benefit of emissions (B_i') and/or attach a lower weight to the consequences of climate change (λ_i). Then an ICA may entail transfers from richer to poorer countries, as the former are willing to compensate the latter in exchange for deeper emissions cuts. Furthermore, if richer countries have more efficient environmental technology than poorer countries, there may be scope for transfers of technology from the former to the latter. But aside from affecting the direction of international transfers within an ICA, we would not expect cross-country asymmetries of this kind to affect the main qualitative insights of our model.

of a country becomes higher as more and more countries collapse. Furthermore, it is direct to verify that Γ_k decreases with k.[41] Thus, the ICA will let the most vulnerable country collapse if and only if $\Gamma_1 > \psi_1$, and in this case, the sequence of country collapses under the ICA will stop when ψ_k rises above Γ_k.

We can now turn to a key question: Will the ICA save some countries that would have collapsed in its absence? Or is it possible that more countries could collapse under the ICA than in its absence? A first point is immediate: if countries fail to coordinate to a pareto-undominated equilibrium and country k would not be saved in the noncooperative setting as a result of this coordination failure, then just as in our common-brink analysis the ICA has a coordination role to play in preventing (in this case, country k's) collapse. With this observation recorded, we resume for the remainder of this section our emphasis on pareto-undominated equilibria in the noncooperative setting, and we focus on comparing the marginal surviving countries under the ICA ($\bar{k}^{ICA}$) and under the noncooperative scenario ($\bar{k}^{N}$).

Recall that $\bar{k}^{ICA}$ is the lowest k such that condition (4.8) is satisfied, while $\bar{k}^{N}$ is the lowest k such that the no-defect conditions (4.4) and (4.5) are both satisfied. First note that the right hand side of (4.8), ψ_k, which is the global average per-capita refugee cost imposed by the collapse of country k, is the population-weighted average of the right hand sides of (4.4) and (4.5), which are respectively the per-capita refugee costs imposed by the collapse of country k on the remaining countries (R_k) and on country k itself (L). On the other hand, as we show in the Appendix, the left hand side of (4.8), Γ_k, is lower than the population-weighted average of the right hand sides of (4.4) and (4.5), G_k and G_k^0. Intuitively, under the ICA the emissions level is symmetric across countries, while under the self-help equilibrium they are asymmetric, with the brink country k choosing zero emissions, and since the gross benefit function $B(c)$ is concave, the average gain from raising emissions is higher in the latter case. This in turn implies that $\bar{k}^{N}$ must be weakly higher than $\bar{k}^{ICA}$, so the set of countries that survive under the ICA is weakly larger than the set of countries that survive under the noncooperative scenario.

It is important to note that the ICA has a potential role to play in saving countries from

[41] To see this, note that

$$\frac{d\Gamma_k}{dk} = \lambda\rho\frac{d\tilde{C}_k}{dk} - B'\left(\frac{\rho\tilde{C}_k}{M}\right)\frac{\rho}{M}\frac{d\tilde{C}_k}{dk} = \frac{\rho}{M}\frac{d\tilde{C}_k}{dk}\left(M\lambda - B'\left(\frac{\rho\tilde{C}_k}{M}\right)\right).$$

Next note that $\frac{d\tilde{C}_k}{dk} > 0$ and $B'\left(\frac{\rho\tilde{C}_k}{M}\right) > M\lambda$ because $\frac{\rho\tilde{C}_k}{M} < \bar{c}^{ICA}$ and $\bar{c}^{ICA}$ is defined by $B'(\bar{c}^{ICA}) = M\lambda$. This implies $\frac{d\Gamma_k}{dk} < 0$.

collapse for two distinct reasons. The first reason was highlighted just above and has to do with the "gains-from-collapse" side of the tradeoff (Γ_k is lower than the average of G_k and G_k^0); the second reason has to do with the "refugee-cost-from-collapse" side of the tradeoff. To see this second reason, suppose as a thought experiment that Γ_k were *equal* to the population-weighted average of G_k and G_k^0, in which case (4.8) would be a side-by-side weighted average of (4.4) and (4.5). Then it is easy to see that the set of surviving countries would *still* be weakly larger under the ICA than under the noncooperative scenario, and strictly larger for a whole parameter region.[42] This is intuitive, since in the noncooperative scenario, if a given country causes the collapse of country k by increasing its own emissions, it exerts a negative refugee externality on all other countries (and this is true also if country k causes its own collapse).

The following proposition summarizes the comparison in the heterogeneous-brink scenario between the set of countries that survive in the noncooperative equilibrium and under the ICA.

Proposition 6. *When the catastrophe points $\tilde{C}_i$ differ across countries, a (weakly) larger subset of countries survive under the ICA than in the noncooperative scenario, but the ICA may still let some countries collapse.*

Recall that (in the absence of coordination failures) a feature of ICAs in our common-brink analysis is their limited useful life. In particular, as recorded in Proposition 2, if the world arrives at the brink under an ICA, then at that point the useful life of the ICA comes to an end. Does this feature extend to the heterogeneous-brink scenario? From our discussion above, it is easy to see that this depends on whether or not the ICA saves some countries relative to the non-cooperative outcome. If it does, clearly the useful life of the ICA extends indefinitely. But if the set of surviving countries is the same under both scenarios, and if the most vulnerable surviving country lives at the brink of collapse in the steady state, then once this country arrives at the brink under the ICA, it will survive at the brink whether or not the ICA remains in place, rendering the ICA useless from this point forward. And if this is the case, then no international transfers will be needed to help this country survive once it arrives at the brink under the ICA, given that the noncooperative outcome is the threat point for ICA negotiations.

[42] We can illustrate this point with an example. Suppose there are only two countries ($M = 2$), so the population of each country is 1/2, and country 1 is at the brink in period t ($C_{t-1} = \tilde{C}_1$). Further suppose, as mentioned in the text, that $\Gamma_1 = (G_1 + G_1^0)/2$, and recall that in this case $\psi_1 = (r + L)/2$. Then (4.4) and (4.5) reduce respectively to $G_1 \leq r$ and $G_1^0 \leq L$, while (4.8) reduces to $G_1 + G_1^0 \leq r + L$. These conditions define paramter regions in (L, r) space, and it is easy to see that the region where (4.4) and (4.5) are both satisfied is a proper subset of the region where (4.8) is satisfied, as claimed in the text.

Finally, recall that we have assumed that countries do not face binding constraints in their ability to make transfers. When we compare the ICA outcome with the social planner's choices that we characterize in the next subsection, this assumption, together with our assumption that ICAs do not face issues of free-riding in participation and compliance, will allow us to focus sharply on the shortcomings of ICAs relative to the social optimum that are due to the inability of future generations to sit at the bargaining table. But it is important to highlight that, if international transfers are limited because of resource constraints, the ICA will have a more limited ability to save countries from collapse than we have characterized here.

Specifically, it is intuitive and can be shown that if international transfers are constrained the ICA lets (weakly) more countries collapse than if international transfers are unlimited; and furthermore, that the ICA is less likely to save a given country if the country faces a more severe constraint on international transfers (because even if it would be efficient to save the country in the presence of unlimited international transfers, the ICA can orchestrate this outcome only if the country has enough resources to compensate the remaining countries and ensure that they receive at least their threat-point payoff when they cut their emissions). This would suggest that smaller countries (like the Maldives) are less likely to be able to look to an ICA to save them from climate catastrophe, because they face more severe resource constraints and hence have less ability to make the substantial transfers to the rest of the world that would be needed under an ICA to achieve this feat.

4.3. The Social Optimum

We next consider the social optimum in the context of the heterogeneous-brink scenario. As international lump-sum transfers are available, within any generation the planner maximizes average per-capita world welfare and uses transfers to redistribute the surplus in each period across countries according to their Pareto weights (which we leave in the background).

As we noted in section 2, given the social discount factor $\hat{\beta} \geq 0$, the objective of the social planner can be written as

$$W = \sum_{t=0}^{\infty} \hat{\beta}^t U_t,$$

where U_t is now given by (4.7). The planner's problem can then be written as:

$$\max \sum_{t=0}^{\infty} \hat{\beta}^t \left[[B(c_t) - \lambda C_t] - \left(\frac{L+r}{M - H_t + 1} \right) \cdot I_t \right]$$

$$\text{s.t. } C_t = (1-\rho) C_{t-1} + \sum_{i=H_t}^{M} \frac{M}{M - H_t + 1} \cdot c_{i,t} \text{ with } C_{-1} = 0$$

$$c_{i,t} \geq 0 \text{ for all } i,t.$$

Again for simplicity we restrict attention to the case where the emissions feasibility constraints $c_{i,t} \geq 0$ are not binding. Given the discontinuities in the payoff functions when the catastrophe point differs across countries, the planner's problem in the heterogeneous-brink scenario is not amenable to a first-order approach as it was in our common-brink scenario, and there is no simple set of optimality conditions that we can write down. But we can establish with direct arguments some qualitative properties of the socially optimal solution.

We focus on a novel feature of the planner's decision in the heterogeneous-brink scenario: How many countries does the planner save from climate catastrophe? We show that a planner's concern for the utility of future generations – as embodied in the social discount factor $\hat{\beta} > 0$ – does not necessarily translate into saving more countries from collapse. In fact, as we next establish, it is possible that a social planner with a higher discount factor will choose a path for emissions that eventually implies a *higher* carbon stock and causes *more* countries to succumb to climate catastrophe than would the same planner with a lower discount factor. And we then consider what this implies for the comparison between the planner's choices and the choices of an ICA regarding the number of countries that are allowed to collapse along the optimal path.

How could it be that a planner with a higher discount factor might choose to allow the carbon stock to rise further and cause more countries to succumb to climate catastrophe than would the same planner with a lower discount factor? The intuition for this surprising result turns out to be as simple as it is illuminating. A planner with a higher discount factor will certainly wish to shift utility toward future generations. The question, though, is how best to do this. Cutting emissions today and lowering the carbon stock that will be inherited by future generations has a direct effect that increases their utility, and this suggests that a planner with a higher discount factor will make emissions choices that lead to a lower carbon stock at any moment in time, and therefore emissions choices that imply (weakly) fewer countries succumbing to climate catastrophe along the optimal path, than would the same planner if it had a lower discount factor. But raising emissions today and crossing the brink of catastrophe

for some country would generate a climate refugee cost that is borne primarily (and under our assumption that this is a one-time cost, solely) by the generation alive today; and if this brink constraint is binding on the emissions choices for future generations that would have been made by the planner in the absence of the brink, then raising emissions today – and incurring today the refugee costs of crossing the brink in order to relax the constraint faced by those alive tomorrow – works indirectly to shift utility toward future generations. As we demonstrate, this indirect effect can dominate the direct effect of the higher carbon stock that would go in the other direction.

More specifically, we show in the Appendix that if for a given social discount factor $\hat{\beta}_1 > 0$ the socially optimal plan does not bring the marginal surviving country to the brink of collapse, then a small increase in the social discount factor, say to $\hat{\beta}_2 > \hat{\beta}_1$, can only reduce (or maintain) the number of countries collapsing along the optimal path. We then show that if instead at the initial discount factor $\hat{\beta}_1$ the marginal surviving country under the optimal plan reaches the brink of catastrophe, then it is possible that the higher discount factor $\hat{\beta}_2$ will lead the country to collapse along the optimal path. In the Appendix we prove:

Lemma 2. *Suppose the catastrophe point $\tilde{C}_i$ differs across countries. If, under the social planner's discount factor $\hat{\beta}_1$, the most vulnerable surviving country along the optimal path stays strictly below the brink, then a higher discount factor $\hat{\beta}_2 > \hat{\beta}_1$ can only lead to the same or fewer countries collapsing along the optimal path. But if under $\hat{\beta}_1$ the most vulnerable surviving country along the optimal path survives at the brink, then it is possible that the higher discount factor $\hat{\beta}_2$ will lead to additional countries collapsing along the equilibrium path.*

And owing to the fact that a social planner with discount factor $\hat{\beta}_1 = 0$ would implement the ICA outcome, we can also state the following:

Proposition 7. *As judged by a social planner with $\hat{\beta} > 0$, the number of countries that collapse under an ICA – and the long-term extent of global warming – can be either too high or too low.*

Proposition 7 confirms that, when countries face heterogeneous climate collapse points, greater concern for the welfare of future generations does not necessarily translate neatly into a reduction in the socially optimal carbon stock at every point in time and an increase in the number of countries that survive along the optimal path.

Finally, it is worth emphasizing the central role played in this finding by our assumption that the refugee costs L and r associated with country collapse are not permanent, but rather are borne primarily by the generation alive at the time of the country's collapse. In our formal model, we have adopted an extreme version of this assumption, namely, that these are one-time costs, but our results would survive as long as these costs are concentrated sufficiently on the current generation. On the other hand, if L and r were instead assumed to reflect constant refugee costs that are incurred in the period of country collapse and every period thereafter, then increasing $\hat{\beta}$ would unambiguously make it (weakly) less likely that a given country k collapses under the social planner's optimal plan. The reason is that (i) the direct effect of increasing the carbon stock inherited by future generations on their utility would be the same as above; and (ii) the indirect effect would be absent, because conditional on being at the brink, if refugee costs were constant and incurred in every period into the infinite future, crossing the brink would *not* shift utility toward future generations. Hence, it is important for Lemma 2 and Proposition 7 that the refugee costs to the world that would be associated with a country's collapse are concentrated sufficiently on the current generation.

5. Intergenerational Altruism

We now extend our analysis to the case in which citizens care about their offspring, that is, we now allow β to be positive. We start by revisiting the common-brink scenario of Section 3, and then we will turn to the heterogeneous-brink scenario of Section 4.

5.1. Common brink

Consider first the noncooperative outcome in the common-brink scenario. Recall that in our setting of successive generations, if agents place positive weight β on their offspring they will maximize the present value of the stream of future utilities with a discount factor β, as defined by the dynastic utility function given in (2.1), even though each agent has only a single offspring. In our infinite-horizon setting, this would give rise to a vast multiplicity of equilibria, including equilibria where players can sustain cooperative behavior by punishing past actions (e.g. trigger-strategy equilibria), or in other words, self-enforcing agreements. But in this paper we want to abstract from issues related to self-enforcing agreements in order to stack the deck in favor of international agreements, and for this reason when characterizing the ICA outcome we have assumed that externally enforced agreements are available. Given this approach, it seems

reasonable when considering noncooperative equilibria in a setting where β is positive to focus on equilibria where current behavior is not conditioned on past actions, and then compare those equilibria with externally-enforced agreements. One way to do this is to focus on a finite-horizon game, because in this case the logic of backward induction unravels self-enforcing agreements. Thus in this section we will consider a version of our game with T periods.[43]

Ideally we would like to consider a game with a large number of periods T, but the characterization of noncooperative outcomes when β is positive turns out to be analytically quite complex, due to the possibility of catastrophe. So our strategy in this section is to solve analytically a two-period version of the game, which will bring about some new dynamic insights, and then turn to numerical methods to obtain results for a larger number of periods.

We therefore consider now the two-period version of our common-brink scenario with $\beta > 0$, focusing on subgame perfect equilibria when characterizing the noncooperative outcome. To streamline the exposition we suppose that there are only two countries, which we label A and B. The formal setup is identical in all other respects to the common-brink scenario we analyzed in Section 3.

A first observation concerns the BAU path of emissions, which we define as the equilibrium path of per-country emissions if the catastrophe threshold $\tilde{C}$ is not binding (i.e. it has no impact on equilibrium emissions), and which we denote by (c_1^{BAU}, c_2^{BAU}). It is easy to see that $c_1^{BAU} < c_2^{BAU}$, since the first generation cares about the next generation, whereas the second generation has no offspring. In what follows we will focus on the more interesting case where the catastrophe threshold $\tilde{C}$ is binding so that the noncooperative equilibrium differs from the BAU emission level, but BAU emissions will still provide an important benchmark.

[43] Two comments are in order. First, in the literature on dynamic games there is another approach to capturing "non-cooperative" behavior, namely focusing on Markov-perfect equilibria of the infinite-horizon game, where current behavior is conditioned only on the current value of the state variable(s) and not on the players' past actions. This approach is effective in infinite-horizon games where it is not possible to credibly punish past actions by using Markov strategies. But in games with global public resources such as ours, where the stock of the resource is the state variable, it has been shown by papers such as Dutta and Sundaram (1993) and Battaglini et al. (2014) that there is a vast multiplicity of Markov-perfect equilibria, including some where players punish past actions indirectly by conditioning current behavior on the current value of the state variable. Thus, in our framework, the idea of characterizing non-cooperative behavior by focusing on Markov-perfect equilibria is not viable. The second comment is that, in some finite-horizon games where the stage game has multiple equilibria, some cooperation can be sustained by specifying that bad actions will be punished by switching to a worse equilibrium of the stage game (see for example Dixit, 1987). In our model, the stage game can have multiple equilibria, for example if players are at the brink of catastrophe, but we will assume that the equilibrium played by the government is not conditional on past actions (and more specifically that they focus on the most efficient symmetric equilibrium). This seems natural given that we want to abstract from self-enforcing agreements.

We proceed by backward induction. The analysis of the subgame ($t = 2$) is straightforward. If the carbon stock inherited from period 1, C_1, is lower than a critical level $\bar{C}_1$, the catastrophe threshold is not binding in period 2 and countries choose their BAU emissions c_2^{BAU}; but if the carbon stock is between $\bar{C}_1$ and $\tilde{C}$, the catastrophe threshold is binding, and the two countries will cut their emissions below c_2^{BAU} in a symmetric way to avoid catastrophe (we can ignore levels of the carbon stock higher than $\tilde{C}$ because they cannot arise in equilibrium).

Next focus on period 1. Here the equilibrium emissions depend crucially on the tightness of the catastrophe threshold $\tilde{C}$. The key step for characterizing the equilibrium emissions is to understand the shape of a country's reaction function. Recall that countries are symmetric, so it suffices to describe country A's reaction function, which is depicted in Figure 4(a).

We draw two locuses in Figure 4(a): the combination of emissions such that $c_1^A + c_1^B = \bar{C}_1$ and the combination of emissions such that $c_1^A + c_1^B = \tilde{C}$ (which we refer to as the Brink1 locus). Both of these locuses are lines with slope -1. Note that in the region left of the $c_1^A + c_1^B = \bar{C}_1$ locus (which we refer to as the No-Brink region) the brink is never reached; in the region between the two locuses (which we refer to as the Brink2 region) the brink is reached at $t = 2$; along the Brink1 locus the brink is reached in the current period ($t = 1$); and in the region to the right of the $c_1^A + c_1^B = \tilde{C}$ locus, catastrophe would occur.

We are now ready to describe the shape of country A's period-1 reaction function. To this end, let us fix $\tilde{C} > 0$ and consider how country A's optimal emissions c_1^A vary as we increase country B's emissions c_1^B from 0 to $\tilde{C}$ (of course we can ignore the case $c_1^B > \tilde{C}$, since it can never be optimal for country B to cause catastrophe). We will first simply describe the shape of the reaction function, and we then explain the reason for the described shape.

If c_1^B is between 0 and some critical level $\bar{c}_1 \geq 0$, the catastrophe constraint does not bind and country A chooses the BAU emission level c_1^{BAU}. Note that the interval $(0, \bar{c}_1)$ will be empty if $\tilde{C}$ is small enough; Figure 4(a) focuses on the case where this interval is nonempty. Note also that this part of the reaction function lies entirely in the No-Brink region. At $c_1^B = \bar{c}_1$, country A's best response jumps up from the BAU level to a level between c_1^{BAU} and c_2^{BAU}, bringing countries into the Brink2 region. As c_1^B increases beyond $\bar{c}_1$, country A's best response decreases with slope flatter than -1 until it reaches the Brink1 locus, and then it runs along that locus (thus it decreases with slope -1) until $c_1^B = \tilde{C}$, at which point country A's best reply is to emit zero.[44]

[44] The part of the reaction function that runs along the Brink1 locus may be empty. For example, if $\rho = 0$

Why does country A's reaction function take on this peculiar shape? While it is intuitive that the reaction function is flat at the BAU level if c_1^B is low and is decreasing if c_1^B is higher, the feature that at some point it jumps above the BAU level and stays above that level for a whole interval of c_1^B is due to an interesting effect that deserves emphasis. We label this the *dynamic free rider (DFR) effect*: anticipating that, if the world reaches the brink in the next period, the burden of cutting emissions to save the world will be shared by both countries, in the current period an individual country has a stronger incentive to raise its emissions, and for this reason the best-response emissions may be above the BAU level.[45]

Another interesting feature of the reaction function is that it has slope flatter than -1 in the Brink2 region. This is because, conditional on the brink being reached at $t = 2$, if country B increases its emissions at $t = 1$, thus forcing country A to reduce its own emissions in order to avoid catastrophe, the optimal way for country A to achieve this reduction is to spread it over both periods for consumption-smoothing reasons.

With the help of Figure 4(a) it is now easy to characterize how the (symmetric) noncooperative equilibrium emissions vary with $\tilde{C}$, given that the equilibrium is determined by the intersection between the reaction function and the diagonal. The key observation is that, as $\tilde{C}$ falls, the reaction function shifts horizontally to the left. In what follows we assume that if there are two equilibria the countries focus on the more efficient one, but the qualitative conclusion would not change in the opposite case. The resulting equilibrium first-period emissions are labeled c_1^e in Figure 4(b).

If $\tilde{C}$ is large enough, the catastrophe constraint has no impact on the equilibrium emissions, so $c_1^e = c_1^{BAU}$. If $\tilde{C}$ lies in the intermediate interval $(\tilde{C}_1, \tilde{C}_2)$, the equilibrium emissions are

and the utility function satisfies the Inada condition ($u(c) \to -\infty$ as $c \to 0$), then reaching the brink at $t = 1$ cannot be an equilibrium for any $\tilde{C} > 0$: emitting a positive amount at $t = 1$ that causes the world to arrive at the brink at $t = 1$ cannot be an equilibrium, because a country would deviate and move some of those emissions to $t = 2$.

[45] The upward jump in the reaction function is due to the fact that, if c_1^B is relatively close to $\bar{c}_1$, country A's payoff function has two local maxima. The first one is c_1^{BAU}, which is the optimal emission subject to (c_1^A, c_1^B) lying in the No-Brink region; this is defined by the first-order condition $u'(c_1^A) - \lambda = \beta\lambda$ and depicted in Figure 4(a) as the dashed line within the No-Brink region. The second local maximum is the optimal emission subject to (c_1^A, c_1^B) lying in the Brink2 region; this is defined by the first-order condition $u'(c_1^A) - \lambda = \beta u'\left(\frac{\tilde{C} - c_1^A - c_1^B}{2}\right)$ and is depicted as the dashed curve within the Brink2 region. Note that the difference between the two first-order condition is that in the former one, the future marginal cost of emissions is the cost of an increase in the future carbon stock at $t = 2$, while in the latter one it is the cost of a (shared) reduction in consumption at $t = 2$. At $c_1^B = \bar{c}_1$ the two local maxima attain the same value, and so at this point the optimal c_1^A jumps from one to the other.

above c_1^{BAU}: this is the reflection of the DFR effect highlighted above.[46] And if $\tilde{C}$ is below $\tilde{C}_1$, the equilibrium emissions are below c_1^{BAU}: here the DFR effect is still at work (because it's still true that the burden of saving the world tomorrow will be shared), but since the catastrophe threshold is tight it is outweighed by the pressure to keep emissions low in both periods, coupled with the need to smooth consumption over time.

The analysis above suggests a number of further questions. How does the DFR effect manifest itself in the noncooperative outcomes, if at all, with a larger number of periods ($T > 2$)? And are there new strategic effects that arise with $T > 2$? Our next step is to employ a numerical approach to examine the game for a larger number of periods so that we can shed light on the answers to these questions. Due to computational constraints, we extend the number of periods to $T = 4$, striking a compromise between achieving a high level of precision with the available computational resources and having a sufficient number of periods to allow for the main interesting effects to arise.

The key results of our numerical analysis are illustrated in Figures 5(a)-5(d). In each figure, the top panel plots the evolution of the global carbon stock C_t over time while the bottom panel plots the evolution of emissions c_t. Noncooperative equilibrium magnitudes are shown as solid red lines, and the BAU magnitudes are shown as dashed blue lines.

For the model parameters described in Figure 5(a), the brink of catastrophe is reached at the end of period 2, and the DFR effect causes noncooperative emissions to rise above their BAU level in period 1. Here, the first generation free rides on the efforts of the second generation in the other country, as it understands that the other country's second generation will reduce emissions to avoid going over the brink in period 2. For the model parameters described in Figure 5(b), the DFR effect causes noncooperative emissions to rise above their BAU level in period 1 but the brink of catastrophe is not reached until the end of period 3. Here, the first generation free rides on the efforts of the next *two* generations in the other country, as both of those generations will reduce emissions to deal with the approaching brink at the end of period 3 (with generation 2 spreading emission cuts over two periods for consumption-smoothing reasons).

For the model parameters described in Figure 5(c), the brink is avoided altogether. But this

[46] Given that with $\beta > 0$ the DFR effect may lead first-period emissions to go above the BAU level, it is natural to ask whether increasing β from zero might paradoxically lead to higher equilibrium emissions. The answer is no: it is easy to show that, even if first-period emissions are above c_1^{BAU}, they can never be above c_2^{BAU}, which recall is the equilibrium emissions level under $\beta = 0$.

is accomplished with the help of the first two generations, who keep their emissions below the BAU level so as to prevent the carbon stock from rising to the point where the DFR effect would kick in for generation 3 and the associated inefficiencies would be incurred. In other words, here the earlier generations reduce emissions as a commitment device to avoid the costs of the DFR effect for later generations. An interesting feature of this equilibrium path is that, even though the brink is not reached in equilibrium, and thus in this sense the catastrophe threshold is not "binding," the mere possibility of catastrophe does affect equilibrium emissions, reducing them (for generation 2) relative to what they would be if catastrophe were not a possibility.

Finally, for the model parameters described in Figure 5(d), the brink is again avoided altogether, but now this is due solely to the efforts of generation 2, whose emissions are kept below the BAU level so as to prevent the carbon stock from rising to the point where the DFR effect would kick in for generation 3. In fact, as Figure 5(d) depicts, for these parameters there is a "second-order" DFR effect that drives the emissions level of generation 1 *above* its BAU level, as generation 1 free-rides on the future efforts of the other country's generation 2 to avoid the DFR effect for generation 3.

Having considered the noncooperative outcome in the common-brink scenario with intergenerational altruism, we next consider the potential role that an ICA can play for improving over the noncooperative equilibrium. Clearly, a key insight from our Section-3 analysis continues to hold with $\beta > 0$: the ICA has a role to play by slowing down the pace of warming while the world is not yet at the brink, but once the world reaches the brink, the ICA no longer has a role to play (beyond possibly helping to solve a coordination failure as we have pointed out in earlier sections). For example, in the model with $T = 4$ analyzed numerically above, if the world reaches the brink under the ICA at the end of period $t = 2$ or $t = 3$, at that point the ICA can be abandoned without any loss. At the same time, our analysis here suggests that the ICA may have an additional role to play relative to the case of no intergenerational altruism: if a dynamic free rider effect is at play in the noncooperative scenario, so that emissions are above the BAU level in the run-up to the brink, the ICA can address this dynamic free riding, in addition to addressing the more standard (static) free riding behavior.

Finally, we revisit the comparison between the ICA solution and the social optimum with intergenerational altruism. In our Section-3 analysis we assumed $\beta = 0$ and allowed the social optimum to place some weight directly on future generations, so that $\hat{\beta} \geq \beta = 0$. Now we focus on the case $\beta > 0$, with the social discount factor again allowed to be weakly higher than β, so

that $\hat{\beta} \geq \beta > 0$.

Recall that the ICA solution coincides with the social optimum in the special case $\hat{\beta} = \beta$, so here we focus on the non-trivial case where $\hat{\beta} > \beta$. In this case, comparing the ICA with the social optimum boils down to comparing the social optimum for two different values of $\hat{\beta}$, a lower level $\hat{\beta}_0 = \beta$ (corresponding to the ICA outcome) and a higher level $\hat{\beta}_1 > \hat{\beta}$ (corresponding to the social optimum). Viewed in this light, the only difference relative to our analysis of Section 3 is that there we focused on the case $\hat{\beta}_0 = 0$, while now we focus on the case $\hat{\beta}_0 > 0$. Clearly for this comparison we are not bound by the same complexities that arise in the noncooperative setting analyzed above, so we can consider an arbitrary number of periods T.

First focus on the impact of raising $\hat{\beta}$ on the optimal emission path c_t. This impact hinges on whether or not the catastrophe constraint is binding. It is not hard to show that: (a) if the catastrophe constraint is not binding for $\hat{\beta} = \hat{\beta}_0$ (and hence is not binding for the higher level $\hat{\beta}_1$ either), raising $\hat{\beta}$ lowers the optimal level of emissions c_t for all generations (with the exception of generation T, who has no offspring); (b) if the catastrophe constraint is binding both for $\hat{\beta}_0$ and $\hat{\beta}_1$, so that for both levels of $\hat{\beta}$ the level of emissions c_t is declining until the brink is reached and then stays constant, raising $\hat{\beta}$ reduces the level of emissions c_t for earlier generations and leads to a more gradual decline in c_t, with the brink being reached later; and (c) if the catastrophe constraint is binding for $\hat{\beta} = \hat{\beta}_0$ but not for $\hat{\beta} = \hat{\beta}_1$, raising $\hat{\beta}$ again lowers emissions for all generations (again with the exception of generation T).

The impact of raising $\hat{\beta}$ on the optimal utility path u_t is more straightforward. It is intuitive and easy to show that raising $\hat{\beta}$ dictates a redistribution of utility from earlier generations to later generations, with the only caveat that the utility of generations who live at the brink under both levels $\hat{\beta}_0$ and $\hat{\beta}_1$ is not affected at all.

The comparison between the ICA solution and the social optimum is an immediate corollary of the observations above. The key qualitative change relative to the case $\beta = 0$ analyzed in Section 3 is that in the case $\beta > 0$, the paths of emissions and utility under the ICA are smoother when the catastrophe constraint is binding on the ICA, with the "brink generation" suffering a less precipitous drop in the level of utility. But we note that, if β is positive but sufficiently small, under the ICA the brink generation will still suffer a larger drop in utility relative to the previous generations. And it remains true that the social optimum will smooth out these paths relative to the ICA case. Thus, at a broad level, the main qualitative insights for the common-brink scenario that we highlighted in the case $\beta = 0$ case continue to hold in

the case $\beta > 0$.

5.2. Heterogenous brinks

We now turn to the question of how the results of our heterogeneous-brink scenario of Section 4 are affected by intergenerational altruism. Mirroring our analysis of the common-brink scenario just above, we consider the two-period version of our heterogeneous-brink scenario with $\beta > 0$, focusing on subgame perfect equilibria when characterizing the noncooperative outcome, and we suppose that there are only two countries, country A (who we now take to be the country more resilient to climate change) and country B (who we take to be the more vulnerable country). The formal setup is identical in all other respects to the heterogenous-brink scenario we analyzed in Section 4. As above, our strategy here is to solve analytically a two-period version of the game while turning to numerical methods to obtain results for a larger number of periods.

A first observation is that, if the internal and external refugee costs (L and R) are infinite, the equilibrium outcome in this scenario is the same as in the common-brink scenario considered just above, since the more resilient countries will view the collapse of a more vulnerable country as a catastrophe also for themselves. And clearly, the same statement applies if L and R are sufficiently large.

Next we turn to the more realistic case where the external refugee cost R is lower than the internal cost L. But before proceeding with this case, we make an important observation. Given $\beta > 0$, if R is sufficiently close to zero the externality caused by the collapse of a country on the surviving countries may be positive. The reason is that, if a country collapses and its citizens migrate, in the following period the world population will be concentrated in a smaller number of countries, who therefore will better internalize the horizontal externalities from emissions, thus leading to a more efficient level of emissions. If R is sufficiently small, this indirect positive "internalization effect" will outweigh the direct negative externality R, thus the *net* externality is positive. And if this is the case, a pure-strategy equilibrium may fail to exist. To see this intuitively in our two-country setup, observe that if the net refugee externality is positive, country A has an incentive to increase its emissions just enough to push country B over the brink. But given country A's emissions, country B's best response is to reduce its emissions just enough to keep the carbon stock exactly at the brink level. Country A could increase its emissions enough to make it infeasible for country B to save itself, but this discrete increase in emissions may be too far from its BAU level to be worth it, and as a consequence

there may be no pure strategy equilibrium.

Given the observation above, and since empirically the negative impact of climate refugees on the receiving countries is arguably of first-order importance, it seems reasonable to assume in what follows that R is not too small relative to β, so that the net refugee externality is non-positive and a pure-strategy equilibrium exists. Treating this as a maintained assumption, we can now turn to the question of whether our main qualitative results of Section 4 continue to hold if β is positive.

Focus first on Proposition 5(i), which states that under some conditions on parameters, and in particular if R is not too large, some countries will collapse on the noncooperative equilibrium path. While we do not allow R to be close to zero when $\beta > 0$ for the reasons discussed above, we can confirm that with $\beta > 0$ there continues to exist a region of parameters (where R is neither too small nor too large) such that some countries collapse on the equilibrium path. This can be confirmed numerically: for the case of two countries and $T = 4$, we checked that there exist parameter values with $\beta > 0$ such that the more vulnerable country collapses in a (pure-strategy) equilibrium. Also the insight of Proposition 5(ii) continues to hold with $\beta > 0$: by the very nature of our model the basic "domino effect" is still present when $\beta > 0$; and the "anti-domino effect" is also still present, in the sense that the refugee externality is more severe if more countries have collapsed in the past.

Focus next on the results of Propositions 6 and 7. Proposition 6 says that the ICA leads to a weakly smaller number of countries collapsing than in the noncooperative scenario. To probe the robustness of this result when $\beta > 0$, recall that the proof of Proposition 6 focuses on the tradeoff that determines the marginal surviving country, and in particular on a comparison across the ICA and the noncooperative equilibrium of the "gains-from-collapse" side of this tradeoff with the "refugee-cost-from-collapse" side of the tradeoff. But the refugee cost from collapse is a one-time cost that is incurred in the period when collapse occurs, so this is independent of β, and our arguments in the discussion of Proposition 6 with regard to this side of the tradeoff still apply when $\beta > 0$. And while the gains-from-collapse side of the tradeoff will be impacted by the level of β, the reason that we describe in the discussion of Proposition 6 as to why the gains from collapse will be smaller in the context of the ICA than in the context of the noncooperative equilibrium are still valid when $\beta > 0$.[47] Hence, with these arguments, it can be

[47] As we describe in the run up to Proposition 6, the reason for why the gains from collapse will be smaller in the context of the ICA than in the context of the noncooperative equilibrium relates to the symmetry of emissions under the ICA versus the asymmetric emissions associated with the self-help equilibrium in the noncooperative

shown that the results of Proposition 6 extend to the two-period version of our heterogeneous-brink scenario with $\beta > 0$. To get some sense of the robustness of this result for a larger number of periods, we turn to our numerical analysis, again for the case of two countries and $T = 4$. We considered a wide range of parameter values (subject to the condition that a pure-strategy equilibrium exists), and in all cases we found that a weakly larger number of countries survive under the ICA than in the noncooperative equilibrium. Regarding Proposition 7, note that this proposition continues to hold with $\beta \geq 0$ for the simple reason that it is a possibility result, and the case $\beta = 0$ is a special case of the more general case $\beta \geq 0$.

Our final observation concerns the dynamic free-rider effect highlighted in the previous subsection. While this effect is sharpest in the common-brink scenario, it is present also in the heterogeneous-brink scenario, but now with a twist, namely that it is asymmetric and tilted toward the more resilient countries. For example, suppose that the world is relatively close to the brink of the weaker country in our two-country setup: anticipating that the weaker country will do whatever it takes to save itself, the stronger country may free-ride on that effort in the previous periods and increase its emissions above the BAU level, while the weaker country emits below the BAU level. We can confirm this asymmetric DFR effect analytically in the two-period, two-country version of the game: it is not hard to show that there is a region of parameters such that at $t = 1$ the strong country emits above the BAU level and the weak country emits below the BAU level, and at $t = 2$ the brink of the weak country is reached but not crossed, with the weak country emitting less than the strong country.

6. Conclusion

The world appears to be facing imminent peril from climate change, as countries are not doing enough to keep the Earth's temperature from rising to catastrophic levels and various attempts at international cooperation have failed. Why is this problem so intractable? Can we expect an 11^{th}-hour solution? Will some countries, or even all, succumb to climate catastrophe on the equilibrium path? In this paper we have addressed these questions through a formal framework that features the possibility of climate catastrophe and emphasizes the role of the international externalities that a country's policies exert on other countries and the intertemporal externalities that current generations exert on future generations. We have examined the interaction between these features and have explored the extent to which international agreements can

setting. This reason is still valid when $\beta > 0$.

mitigate the problem of climate change in their presence. Our analysis delivers novel insights on the role that international climate agreements can be expected to play in addressing climate change, and it points to important limitations on what such agreements can achieve, even under the best of circumstances.

We have adopted many strong assumptions to carry out our analysis. For example, throughout we have ignored potentially important "domestic" problems which might also frustrate attempts to address climate change, such as the political power of the fossil fuel industry. Putting such issues aside has the advantage of producing a modeling framework that is capable of generating novel insights with maximum clarity. But it also carries the risk of abstraction from important features that should not be ignored in a more complete analysis of the issues. In this light, it seems appropriate to conclude with a discussion of a number of further extensions of our framework and analysis which seem especially salient. Our purpose here is not to provide a full analytical treatment of these extensions. Rather, we keep our discussion brief, and focus only on a few key points.

Uncertainty and/or heterogeneous beliefs We have abstracted from uncertainty, but of course uncertainty is an important feature of climate change and the challenge that it poses for the world.[48] Moreover, and relatedly, beliefs about climate change are heterogeneous, with some countries and some groups within countries firmly believing that a climate catastrophe will occur if the world continues along its BAU path, while others are skeptical of such claims or even deny outright that the threat of climate catastrophes are real. How would these features affect our analysis?

To explore this question, we focus on uncertainty over the level of the catastrophic global carbon stock. In our common-brink scenario, a very stylized way to introduce such uncertainty is to assume that with probability κ the catastrophe threshold for the carbon stock is $\tilde{C}$, and with probability $(1-\kappa)$ this threshold is infinite. With $\tilde{C}$ satisfying Assumption 1, this would imply that with probability κ a climate catastrophe will occur if the world continues along its BAU path and with probability $(1-\kappa)$ there is no climate catastrophe to worry about. And similarly in our heterogeneous-brink scenario, we can assume that with probability κ_i the critical level of the carbon stock for country i is $\tilde{C}_i$ and with probability $(1-\kappa_i)$ it is infinite.

[48] Uncertainty regarding the precise location of thresholds and tipping points is widely emphasized in the climate literature (see, for example, Lenton et al., 2008, 2019, and Rockstrom et al., 2009).

With these assumptions, it is straightforward to show that introducing the implied uncertainty about the location of the critical carbon thresholds into our analysis does not change our basic findings. In the common-brink scenario, as long as the cost of exceeding $\tilde{C}$ continues to be infinite, any strictly positive probability κ that the critical carbon stock is $\tilde{C}$ rather than infinite will ensure that countries will not exceed the level $\tilde{C}$ in the noncooperative equilibrium, just as in our common-brink analysis of section 3.[49] And similarly, in the heterogeneous-brink scenario, country i will avoid collapse with certainty in the noncooperative equilibrium if $\kappa_i R_i$ and $\kappa_i L$ are above some thresholds and will collapse with probability κ_i otherwise, while under the ICA and socially optimal choices country i will avoid collapse with certainty if $\kappa_i \psi_i$ is above some threshold and will collapse with probability κ_i otherwise. In this setting, our main findings of section 4 continue to hold, and in particular, the set of surviving countries under the ICA is weakly larger than under the noncooperative scenario, but may be larger or smaller than under the social optimum.

We can also consider the possibility that beliefs are heterogeneous across countries in our common-brink scenario with a simple reinterpretation of the parameter κ introduced just above, by assuming that $X \in \{\tilde{C}, \infty\}$ is the true level of the critical carbon stock and that κ now represents the fraction of countries that believe the critical carbon stock is $\tilde{C}$, with the remaining fraction $(1 - \kappa)$ of countries believing that the critical carbon stock is infinite and hence that there is no climate catastrophe to worry about. And similarly for the heterogeneous-brink scenario, we can capture the possibility of heterogeneous beliefs by assuming that $X_k \in \{\tilde{C}_k, \infty\}$ is the true level of the critical carbon stock for country k and that κ_k now represents the fraction of countries that believe the critical carbon stock for country k is $\tilde{C}_k$, with the remaining fraction $(1 - \kappa_k)$ of countries believing that the critical carbon stock for country k is infinite.

Interestingly, this reinterpretation suggests that heterogeneous beliefs about the risks posed by climate change could potentially be more devastating to the world than uncertainty about the position of the critical thresholds. For example, with heterogeneous beliefs it is easy to see that there is an important new possibility in the common-brink scenario: if $X = \tilde{C}$ so

[49] Note that the expected cost of exceeding $\tilde{C}$ is infinite ($\kappa \times \infty + (1-\kappa) \times \lambda = \infty$), so if agents are risk-neutral their expected utility for $C > \tilde{C}$ is minus infinity, and of course if agents are risk-averse this conclusion is strengthened. Also note that, if the loss from exceeding the threshold $\tilde{C}$ is very high but finite (say $\bar{L}$), and $\tilde{C}$ is a random variable with a bounded support, then the expected loss will be continuous but rising very steeply for C in the support of $\tilde{C}$. In this case, fixing the distribution of $\tilde{C}$, as $\bar{L}$ goes to infinity the expected loss function converges to the one we assumed, and we conjecture that the results would then be approximately the same as those of our common-brink scenario.

that the true level of the critical carbon stock is $\tilde{C}$ and if κ is sufficiently small so that the fraction of climate skeptics and deniers in the world is sufficiently high, then it is possible that in the noncooperative equilibrium and contrary to our common-brink scenario of section 3 the world will trigger a climate catastrophe, because the climate skeptics and deniers are sufficiently prevalent in the world to preclude the possibility that the climate believers of the world could do enough on their own to avoid the catastrophe, much as can happen for the most vulnerable countries who face a climate crisis alone in our heterogeneous-brink scenario of section 4. A further implication is then that in the presence of heterogeneous beliefs a potential new role for ICAs could also arise in the common-brink scenario, namely, that through their reductions in emissions, ICAs might help keep the global carbon stock below $\tilde{C}$ and thereby help the world avoid a climate catastrophe that would occur on the noncooperative equilibrium path. Similar possibilities can arise in the heterogeneous-brink scenario when beliefs are heterogeneous.[50]

Lags We have assumed that each generation incurs the warming generated by its own emissions, and that crossing the critical carbon-stock threshold will subject the generation that crosses the threshold to a climate catastrophe. If we think of a generation as representing a span of 20 to 30 years, then our analysis can accommodate lags on the order of a decade or two in the process by which a rising carbon stock leads to warming temperatures, and our findings apply without modification in the presence of such lags.

But the presence of longer lags in the effects of emissions may be especially relevant in the context of various "tipping point" phenomena, whereby if the global carbon stock crosses a certain threshold, an inevitable and irreversible process is set in motion that leads to climate catastrophe several generations into the future.[51] Clearly, in the absence of intergenerational altruism such a generation-spanning lag could cause a climate catastrophe to occur in the noncooperative equilibrium of the common-brink scenario, contrary to our findings in section 3. Still, if there is any intergenerational altruism and provided the cost of a climate catastrophe in the common-brink scenario is infinite, then there will be no climate catastrophe in the non-

[50] For example, if a sufficient fraction of the rest of the world does not believe that country k has reached the brink of catastrophe at $\tilde{C}_k$ when in fact country k really is on the brink of collapse, then country k may not be able to acquire the help from the world that it needs to avoid collapse, because too few countries believe that they would face refugee externalities from country k if they don't step up their efforts to reduce emissions and keep the global carbon stock from exceeding $\tilde{C}_k$.

[51] On the possibility of climate tipping points and the likely intergenerational lags associated with them, see, for example, Lenton et al (2019).

cooperative equilibrium and our earlier analysis will apply in the presence of intergenerational lags of this nature, much as is true with uncertainty about the position of the critical climate stock as discussed above. And with finite costs of a climate catastrophe, the same statement applies provided that the degree of intergenerational altruism is sufficiently high.

Introducing such lags into our heterogeneous-brink scenario of section 4 would have similar effects with regard to individual countries, increasing the likelihood of country-level climate catastrophes under the noncooperative scenario, the ICA and the social optimum. Still, it is not clear that there would be an impact on the relative outcomes across these three scenarios, and hence it is not clear that there would be new implications for the comparison between noncooperative, ICA and socially optimal outcomes that are our focus.

The possibility of a future technological fix Our analysis abstracts from technological change, but it is important to consider the possibility that a technological fix to the climate problem might arrive at some point in the future. To that end, we briefly consider here the possibility that a technological breakthrough might occur, producing a "silver-bullet" technology that solves the climate problem once and for all, as in Besley and Dixit (2019).

To fix ideas, we imagine that each country takes an iid technology draw in each period, and if a draw is successful a silver-bullet technology is discovered, in which case the catastrophe threshold $\tilde{C}$ jumps to infinity forever thereafter. It is easy to see that the possibility of a silver-bullet discovery would have no impact on our findings in the common-brink scenario of section 3, beyond the obvious impact that if the silver-bullet technology is found then all countries would revert to BAU emissions levels thereafter. But it would have a number of interesting implications within the heterogeneous-brink scenario of section 4.

First, in the heterogeneous-brink scenario there would now be a novel benefit of slowing down the march to the brink for any country, in the sense that with more time before a country reaches the brink there is now a greater chance that the silver-bullet technology will appear and the country will be saved from the brink by a technological fix. Moreover, there could be a new international externality associated with a given country's collapse: for example, if R&D also involves substantial sunk "lab" investments that facilitate the possibility of technology draws and that cannot not be transferred to other countries and would therefore be lost when a country succumbs to a climate catastrophe, then the probability of the arrival of a technological fix at some point in the future is diminished with each country that collapses and gives up its ability

to take a technology draw, making all surviving countries worse off in expectation with each collapse. In the heterogeneous-brink scenario this would make collapse along the equilibrium path less likely for the social optimum, though again it is not clear that there would be an impact on the relative outcomes across the noncooperative, ICA and socially optimal scenarios.[52]

Finally, if it were allowed that investment in technology is a choice that countries make, with increases in investment leading to a greater chance that the technology draw will result in the silver bullet, then these investments themselves would be a natural policy for ICAs to cover, in addition to the emissions choices that we have focused on, since such investments would entail obvious positive externalities. We leave an exploration of this possibility to future research.

Investment in adaptation We have focused on mitigation (i.e., reduced emissions) as a response to the climate problem, but an alternative and possibly complementary response is to invest in adaptation (e.g., building sea walls). It is interesting to consider two possible interpretations of investments in adaptation within our modeling framework. One possibility is that adaptation lowers the cost that a country experiences from moderate levels of warming, related to our parameter λ; this could be captured by allowing each country i to have its own λ_i which it could lower at a cost. A second possibility is that adaptation raises the catastrophe threshold $\tilde{C}_i$ for country i, and here country i could raise $\tilde{C}_i$ at a cost. Under the first interpretation, the possibility of adaptation could be added to either our common-brink scenario or our heterogeneous-brink scenario; to consider the second interpretation it is only the heterogeneous-brink scenario that is relevant.

While mitigation generates a positive externality across countries and across generations and is hence to be generally encouraged under the ICA and in the social optimum relative to the noncooperative outcome, adaptation may generate a negative externality. Under the first interpretation where adaptation reduces λ_i, the externality is unambiguously negative, because country i's noncooperative choice of emissions in both the common-brink and heterogeneous-brink scenarios will be higher when it invests in adaptation and lowers λ_i. Clearly, then, the ICA and the social planner will require that countries increase mitigation but, under this first interpretation, *decrease* adaptation relative to noncooperative choices. Under the second inter-

[52] We have highlighted the case of a negative innovation externality when a country collapses, which seems plausible. But the externality could be positive if the country's collapse led its emigrating scientists to join more productive labs in other countries, as might happen if the collapsing country were relatively poor and there were migration frictions that otherwise prevented the efficient matching of scientist with labs across the world.

pretation, where country i's investments in adaptation raise $\tilde{C}_i$, the international externality may be positive or negative. In particular, if $\tilde{C}_i$ rises but stays below $\tilde{C}_{i+1}$, countries $i+1$ and higher may benefit, to the extent that this makes it less likely that country i collapses on the equilibrium path and hence imposes climate refugee costs on other countries. But if $\tilde{C}_i$ rises above $\tilde{C}_{i+1}$, country $i+1$ may be worse off, to the extent that it will now have to face the brink of collapse before country i. Thus the implications for the treatment of adaptation in ICAs and the social optimum relative to noncooperative levels may be different across these two interpretations.[53]

Adaptation under the second interpretation also suggests a reason that the brink levels in the heterogeneous-brink scenario might be positively correlated with country wealth and income: even abstracting from natural variation across countries in the susceptibility to climate change due to geography, if richer countries are more able to invest in adaptation than poorer countries, then in the noncooperative equilibrium it is likely to be the poorer countries who are most vulnerable to the greatest costs of climate change, and the ones most likely to suffer collapse from a climate catastrophe. How ICAs and the social optimum would alter these distributional impacts of climate change from what would transpire in the noncooperative equilibrium is an important question that we leave for future research.

Finally, two other types of investment may have interesting implications: investments that increase the natural atmospheric regeneration rate ρ, such as reforestation; and investments in cleaner processes and products, which in our model would shift up the schedule $B(c)$ for the whole world (to the extent that such innovations spill over at least partially across countries), by making it possible to achieve the same utility with a lower level of emissions. Intuitively, these kinds of investment should have similar (positive) international externalities as mitigation policies, and thus should be encouraged by both ICAs and the social optimum.

Trade We have abstracted from international trade in our formal analysis, but the links between climate issues and trade policy are central to the debate over the appropriate response to

[53] There is also a third interpretation of adaptation that would be interesting to consider in our heterogeneous-brink scenario: investing in infrastructure to reduce the cost of receiving climate refugees when other countries collapse, or to keep refugees out of the country (e.g. border walls). Such investments by a country i would decrease the cost incurred by country i if another country collapses, but may increase the cost incurred by *other* surviving countries that receive refugees, and may increase the cost incurred by the climate refugees themselves (L), thus the implications of such investments and whether they would be encouraged or discouraged by an ICA or the social optimum is an interesting and open question.

global warming (see Esty, 1994, for an early expression of these links, and see Nordhaus 2015, more recently). In our heterogeneous-brink scenario, we have emphasized climate refugees as an important example of the international externality that is imposed on surviving countries when a country succumbs to a climate catastrophe. And as we have shown, the externalities that a collapsing country would impose on the remaining countries plays an important role in determining – under both the noncooperative emissions choices and also the ICA and socially optimal choices – the path of emissions and whether or not countries collapse along the equilibrium path under these three scenarios.

But another potentially important externality associated with a country's collapse is the loss of the gains from trade with the collapsing country that the remaining countries may suffer. Viewed from this lens, it is direct to see that our model implies a link between trade policy and the climate issues we study: the bigger the gains from trade, the larger will be the negative externalities imposed on surviving countries when a country collapses due to climate change, and by the logic of our model the fewer countries are likely to collapse along the equilibrium path. This suggests in turn that effective cooperation on trade issues, to the extent that such cooperation enhances the size of the gains from trade, can by itself help to reduce the most extreme costs of climate change.

7. Appendix

Proof of Proposition 3

The Lagrangian associated with the planner's problem is:

$$L = \sum_{t=0}^{\infty} \left\{ \hat{\beta}^t [B(c_t) - \lambda C_t] + \xi_t \left[C_t - (1-\rho) C_{t-1} - M c_t \right] + \phi_t (C_t - \tilde{C}) \right\} \tag{7.1}$$

where ξ_t and ϕ_t are Lagrange multipliers. Differentiating (7.1) with respect to c_s yields the first-order condition

$$\frac{\partial L}{\partial c_s} = \hat{\beta}^s B'(c_s) - \xi_s = 0. \tag{7.2}$$

And differentiating (7.1) with respect to C_s yields the first-order condition

$$\frac{\partial L}{\partial C_s} = -\hat{\beta}^s M \lambda + \xi_s - (1-\rho)\xi_{s+1} + \phi_s = 0 \tag{7.3}$$

where we use the fact that each C_s enters two terms of (7.1), the $t = s$ term and the $t = s + 1$ term. Finally, solving (7.2) for ξ_s, substituting into (7.3) and converting s to t, yields

$$-M\lambda + B'(c_t) - (1-\rho)\hat{\beta}B'(c_{t+1}) + \hat{\beta}^{-t}\phi_t = 0. \tag{7.4}$$

The transversality condition is non-standard and requires some care, so we address it below.

To proceed, we will follow a guess-and-verify approach. There are two cases to consider, depending on whether or not the brink constraint $C_t \leq \tilde{C}$ binds for any t.

Case 1: the brink is never reached.

We first suppose that the brink constraint never binds, so we set $\phi_t = 0$ for all t in (7.4).

Note that c_t enters equation (7.4) only through $B'(c_t)$, so we can let $X_t \equiv B'(c_t)$ and treat X_t as the unknown rather than c_t, keeping in mind that X_t is decreasing in c_t. We can thus rewrite (7.4) as the first-order linear difference equation

$$-M\lambda + X_t - (1-\rho)\hat{\beta}X_{t+1} = 0. \tag{7.5}$$

The solutions to (7.5) are characterized by

$$X_t = \frac{K}{\hat{\beta}^t(1-\rho)^t} + \frac{M\lambda}{1-\hat{\beta}(1-\rho)} \tag{7.6}$$

where K is an arbitrary constant. The expression in (7.6) defines a family of curves, one of which is constant (for $K = 0$), while others are increasing and convex (for $K > 0$) and still others are decreasing and concave (for $K < 0$). For future reference, we write the constant solution to (7.6) when $K = 0$ as

$$X_t = \frac{M\lambda}{1-\hat{\beta}(1-\rho)} \equiv X. \tag{7.7}$$

We now argue that only the constant solution described by (7.7) satisfies the first-order conditions (7.2) and (7.3). To make this argument, we consider the finite-T problem and take the limit of the solution as $T \to \infty$.

In the finite-T problem, X_T must satisfy the first-order condition $-M\lambda + X_T = 0$, which follows from (7.5). This determines the transversality condition for the finite-T problem:

$$X_T = M\lambda. \tag{7.8}$$

Note that, since $M\lambda < X$, the curve in (7.6) that satisfies (7.8) must have $\frac{K}{\hat{\beta}^T(1-\rho)^T} < 0$ and hence $K < 0$. This establishes that in the finite-T problem, the optimum path for X_t is not the constant solution described by (7.7), but one of the decreasing paths.

Now consider the limit as $T \to \infty$. As T increases, the curve in (7.6) that satisfies (7.8) gets closer and closer to the constant solution described by (7.7). Indeed, as $T \to \infty$ the solution converges pointwise to (7.7).

Thus our candidate solution for Case 1 is the constant solution $X_t = \bar{X}$, and using $X_t \equiv B'(c_t)$, the associated level of emissions for a representative country and for every generation, which we denote by $\bar{c}^S$, is defined by $B'(\bar{c}^S) = \frac{M\lambda}{1-\hat{\beta}(1-\rho)}$, implying

$$\bar{c}^S = B'^{-1}\left(\frac{M\lambda}{1-\hat{\beta}(1-\rho)}\right) \tag{7.9}$$

This is the optimum if the implied carbon stock never reaches $\tilde{C}$. It is easy to see that, if the emissions level is $\bar{c}^S$ per country, the carbon stock increases in a concave way and converges to the steady state level $\frac{M}{\rho}\bar{c}^S \equiv C^S$, hence the condition for $\bar{c}^S$ to be the solution is

$$\tilde{C} \geq C^S \tag{7.10}$$

where note from their definitions that $C^S < C^N$ so both Assumption 1 and (7.10) will be satisfied if $\tilde{C} \in [C^S, C^N)$.

Finally note that the global stock of carbon, denoted $\bar{C}^S_t$, evolves in Case 1 according to the difference equation

$$\bar{C}^S_t = (1-\rho)\bar{C}^S_{t-1} + M\bar{c}^S \quad \text{with } \bar{C}^S_{-1} = 0.$$

Case 2: the brink is reached in finite time

Now suppose that the critical level of the carbon stock $\tilde{C}$ is below the threshold level C^S so that (7.10) is violated and instead we have

$$\tilde{C} < C^S. \tag{7.11}$$

In this case our candidate Case-1 solution (7.9) does not work, and we need to proceed to the second guess where the brink constraint $C_t \leq \tilde{C}$ binds from some $\tilde{t}^S$ onward.

For $t \geq \tilde{t}^S$, under this guess C_t stays constant at the threshold level $\tilde{C}$, hence c_t must be set at the replacement rate dictated by the natural rate of atmospheric regeneration given by

$$c_t = \frac{\rho\tilde{C}}{M} = \hat{c}^N \quad \text{for } t \geq \tilde{t}^S. \tag{7.12}$$

For $t < \tilde{t}^S$, the guess is that the brink constraint does not bind, so $\phi_t = 0$, and hence we arrive at the same system of first-order difference equations as (7.5), which yields the family of

curves (7.6). Given $\tilde{t}^S$, we pick the solution (i.e., pick K) by imposing the first-order condition (7.5) at $t = \tilde{t}^S$:

$$-M\lambda + X_{\tilde{t}^S} - (1-\rho)\hat{\beta}\hat{X} = 0 \tag{7.13}$$

where $\hat{X} \equiv B'(\hat{c}^N)$. Again ignoring integer constraints, this requires continuity of X_t, and therefore of c_t.[54] But given (7.11) we have that $\bar{c}^S > \frac{\rho\tilde{C}}{M} = \hat{c}^N$. And recalling that $\bar{c}^S$ is defined by the constant solution to (7.6) with $K = 0$ so that $X_t = \bar{X}$, this implies that the socially optimal path of c_t for $t \leq \tilde{t}^S$, which we denote by $\hat{c}_t^S$, must be defined by a solution to (7.6) with $K > 0$ so that $X_t > \bar{X}$. It then follows from (7.6) together with (3.10) that $\hat{c}_t^S$ begins at $t = 0$ at a level that is strictly below $\bar{c}^{ICA}$, is decreasing, and hits $\hat{c}^S$ at $\tilde{t}^S$.[55]

Finally, to determine $\tilde{t}^S$, we use the condition that the path of C_t implied by the path of emissions $\hat{c}_t^S$, which we denote $\hat{C}_t^S$, reaches $\tilde{C}$ at $\tilde{t}^S$. The path $\hat{C}_t^S$ is the solution to the difference equation

$$\hat{C}_t^S = (1-\rho)\hat{C}_{t-1}^S + M\hat{c}_t^S \quad \text{with } \hat{C}_{-1}^S = 0. \tag{7.14}$$

Thus $\tilde{t}^S$ is defined using (7.14) and $\hat{C}_{\tilde{t}^S}^S = \tilde{C}$. Using this condition and the analogous condition (3.11) that defines $\tilde{t}^{ICA}$ as well as the properties of $\hat{c}_t^S$ described above, it is direct to confirm that $\tilde{t}^{ICA} < \tilde{t}^S$.[56]

We may conclude that in Case 2, the socially optimal emissions for generation t in a representative country are given by $c_t^S = \hat{c}_t^S$ for $t < \tilde{t}^S$ and $c_t^S = \hat{c}^N$ for $t \geq \tilde{t}^S$.

Proof of Proposition 6

Recall that the marginal surviving country under the noncooperative equilibrium is the

[54] If we take the integer constraint into account, there will (generically) be a period (say $\tilde{t}^{FB} - 1$) where X_t is between $\hat{X}$ and the level defined by (7.13).

[55] Depending on the functional form of B (and in particular on its third derivative), the implied path of $\hat{c}_t^S$ for $t \leq \tilde{t}^S$ may be concave or convex. For example if B is quadratic, the path is concave, but if B is logarithmic the path is convex.

[56] One might wonder whether there is another potential candidate solution: among the paths that satisfy (7.6), is there one such that the implied carbon stock C_t approaches $\tilde{C}$ as $t \to \infty$, and might this be the optimum? The answer is no. It is easy to show that there is only one solution of (7.6) such that the associated path of C_t converges to a strictly positive level, and that is the $K = 0$ solution, with the associated carbon stock converging to $\bar{C} = \frac{M\bar{c}}{\rho} > \tilde{C}$. For all solutions with $K > 0$, the path of X_t diverges to infinity, thus the path of c_t goes to zero, and hence also C_t converges to zero.

lowest k that satisfies

$$G_k \equiv \left[B(\bar{c}_k^N) - \lambda\left(\tilde{C}_k + \frac{M}{M-k+1}\cdot\left(\bar{c}_k^N - \frac{\rho\tilde{C}_k}{M-k}\right)\right)\right] - \left[B\left(\frac{\rho\tilde{C}_k}{M-k}\right) - \lambda\tilde{C}_k\right] \leq \frac{r}{M-k}$$

$$G_k^0 \equiv \left[B(\bar{c}_k^N) - \lambda\left(\tilde{C}_k + \frac{M}{M-k+1}\cdot\bar{c}_k^N\right)\right] - \left[B(0) - \lambda\tilde{C}_k\right] \leq L$$

We now argue that if the above conditions are satisfied, also the following condition is satisfied, so country k will survive under the ICA:

$$\Gamma_k \equiv \left[B(\bar{c}^{ICA}) - \lambda\left((1-\rho)\tilde{C}_k + M\bar{c}^{ICA}\right)\right] - \left[B\left(\frac{\rho\tilde{C}_k}{M}\right) - \lambda\tilde{C}_k\right] \leq \frac{L+r}{M-k+1} \equiv \psi_k$$

Noting that ψ_k is the population-weighted average of $\frac{r}{M-k}$ and L (where the weights are respectively $\frac{M-k}{M-k+1}$ and $\frac{1}{M-k+1}$), all we need to show is that Γ_k is no higher than the population-weighted average of G_k and G_k^0, that is:

$$\Gamma_k \leq \frac{M-k}{M-k+1}G_k + \frac{1}{M-k+1}G_k^0$$

or equivalently

$$\begin{aligned} & B(\bar{c}^{ICA}) - B\left(\frac{\rho\tilde{C}_k}{M}\right) - \lambda\left(M\bar{c}^{ICA} - \rho\tilde{C}_k\right) \\ \leq\ & \frac{M-k}{M-k+1}\left[B(\bar{c}_k^N) - B\left(\frac{\rho\tilde{C}_k}{M-k}\right) - \frac{\lambda M}{M-k+1}\left(\bar{c}_k^N - \frac{\rho\tilde{C}_k}{M-k}\right)\right] \\ & + \frac{1}{M-k+1}\left[B(\bar{c}_k^N) - B(0) - \frac{\lambda M}{M-k+1}\bar{c}_k^N\right] \end{aligned}$$

Letting $\alpha \equiv \frac{M-k}{M-k+1}$ and $v(c) \equiv B(c) - \lambda\frac{M}{M-k+1}c$, and rearranging, we need to show

$$v(\bar{c}_k^N) - v(\bar{c}^{ICA}) + \lambda\frac{M-k}{M-k+1}(M\bar{c}^{ICA} - \rho\tilde{C}_k) + \left[\alpha v\left(\frac{\rho\tilde{C}_k}{M-k}\right) + (1-\alpha)v(0) - v\left(\frac{\rho\tilde{C}_k}{M}\right)\right] \geq 0$$

This condition is satisfied because: (i) $v(c)$ is maximized by $\bar{c}_k^N$ (recalling that the population of each country at this stage is $\frac{M}{M-k+1}$), so $v(\bar{c}_k^N) > v(\bar{c}^{ICA})$; (ii) concavity of $v(c)$ implies that the square parenthesis is positive; and (iii) for country k to reach the brink under the ICA, it must be $M\bar{c}^{ICA} > \rho\tilde{C}_k$. The claim follows.

Proof of Proposition 7

Let $\bar{k}^S$ denote the marginal surviving country under the optimal plan given discount factor $\hat{\beta}_1$. Suppose first that the planner's discount factor is given by $\hat{\beta}_1 > 0$ and that country $\bar{k}^S$ is not brought to the brink of collapse. This implies that under the optimal plan, the steady state level of the carbon stock remains strictly below $\tilde{C}_{\bar{k}^S}$. If $\hat{\beta}$ is then increased from $\hat{\beta}_1$ to $\hat{\beta}_2$, the optimal carbon path will be adjusted to shift utility to future generations. But if $\tilde{C}_{\bar{k}^S}$ did not impose a binding constraint for any generation under the optimal plan with discount factor $\hat{\beta}_1$, then increasing the carbon stock to exceed the brink will not relax any constraint for future generations, and will instead only leave future generations with a higher inherited carbon stock, which by itself reduces their utility. So in this case, the only way to shift utility to the future when $\hat{\beta}$ increases from $\hat{\beta}_1$ to $\hat{\beta}_2$ is to reduce the carbon stock from the initial level, and that ensures that a higher discount factor $\hat{\beta}_2 > \hat{\beta}_1$ can only lead to the same or fewer countries collapsing along the optimal path.

Now suppose that the optimal plan under the discount factor $\hat{\beta}_1$ brings the marginal surviving country $\bar{k}^S$ to the brink of collapse, implying that under the optimal plan the steady state level of the carbon stock reaches $\tilde{C}_{\bar{k}^S}$ and remains there forever. We wish to show that it is possible in this case that the higher discount factor $\hat{\beta}_2$ will lead country $\bar{k}^S$ to collapse along the optimal path.

To see that this is possible, note first that for a planner with discount factor $\hat{\beta}_1$, the payoff (expressed in per-capita terms) from remaining at the brink $\tilde{C}_{\bar{k}^S}$ forever is

$$\frac{1}{1-\hat{\beta}_1}\left[B\left(\frac{\rho\tilde{C}_{\bar{k}^S}}{M-\bar{k}^S+1}\right)-\lambda\tilde{C}_{\bar{k}^S}\right], \tag{7.15}$$

while the payoff (expressed in per-capita terms) from going over the brink today is

$$B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1))-\lambda\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)-\frac{L+r}{M-\bar{k}^S+1}+\hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1,\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)), \tag{7.16}$$

where $\bar{c}_{\bar{k}^S}(\hat{\beta}_1)$ is the optimal emissions level conditional on going over the brink $\tilde{C}_{\bar{k}^S}$ for the period where the brink is exceeded (i.e., today), $\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) \equiv (1-\rho)\tilde{C}_{\bar{k}^S} + M\bar{c}_{\bar{k}^S}(\hat{\beta}_1)$ is the level of the carbon stock at the beginning of the next period implied by $\bar{c}_{\bar{k}^S}(\hat{\beta}_1)$, and $V_{\bar{k}^S}(\hat{\beta}_1,\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1))$ is the value function (expressed in per-capita terms) beginning in the period after the the brink $\tilde{C}_{\bar{k}^S}$ is exceeded (i.e., beginning tomorrow) when the initial carbon stock is $\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)$.

To demonstrate this possibility result, it is convenient to suppose that the values of L and r are such that under the optimal plan for $\hat{\beta}_1$ the planner is *indifferent* between remaining at

the brink $\tilde{C}_{\bar{k}^S}$ and exceeding the brink. Using (7.15) and (7.16) we then have with this choice of L and r the initial indifference condition

$$B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)) - \frac{1}{1-\hat{\beta}_1}\left[B\left(\frac{\rho \tilde{C}_{\bar{k}^S}}{M-\bar{k}^S+1}\right) - \lambda \tilde{C}_{\bar{k}^S}\right] = \frac{L+r}{M-\bar{k}^S+1}. \quad (7.17)$$

With L and r fixed at their initial levels, the right-hand side of (7.17) is independent of $\hat{\beta}$, so we want to show that the left-hand side of (7.17) can be increasing in $\hat{\beta}$. The derivative of the left-hand side of (7.17) with respect to $\hat{\beta}$ can be written as

$$\frac{1}{1-\hat{\beta}}\left(\frac{L+r}{M-\bar{k}^S+1} + \frac{d}{d\hat{\beta}}\left((1-\hat{\beta})\left[B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1))\right]\right)\right), \quad (7.18)$$

where we have used the initial indifference condition (7.17). If we can show that the expression in (7.18) can be strictly positive, then we will have shown that a small increase in $\hat{\beta}$ starting from $\hat{\beta}_1$ can lead to a strict preference for exceeding the brink $\tilde{C}_{\bar{k}^S}$, causing country $\bar{k}^S$ to collapse and increasing the number of countries that succumb to a catastrophe along the optimal path.

The expression in (7.18) is strictly positive if and only if

$$\frac{L+r}{M-\bar{k}^S+1} > -\frac{d}{d\hat{\beta}}\left((1-\hat{\beta})\left[B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1))\right]\right). \quad (7.19)$$

The left-hand side of (7.19) is strictly positive. Hence, if we can find model parameters such that the right-hand side is equal to zero, we will have established the possibility result. But this is straightforward. Suppose, for example, that $\lambda = 0$ and that $\tilde{C}_{\bar{k}^S+1}$ is sufficiently high so that $\tilde{C}_{\bar{k}^S+1}$ will never bind along the optimal path, even if $\tilde{C}_{\bar{k}^S}$ were exceeded. Then conditional on exceeding $\tilde{C}_{\bar{k}^S}$, the utility of each subsequent generation would be a constant $\bar{V}$ (because consumption would be a constant, and the rising carbon stock would cause no disutility when $\lambda = 0$ and $\tilde{C}_{\bar{k}^S+1}$ is sufficiently high so that $\tilde{C}_{\bar{k}^S+1}$ never binds), and hence

$$B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)) = \frac{1}{1-\hat{\beta}_1}\bar{V},$$

implying

$$\frac{d}{d\hat{\beta}}\left((1-\hat{\beta})\left[B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1))\right]\right) = \frac{d\bar{V}}{d\hat{\beta}} = 0.$$

Hence, if $\lambda = 0$ and $\tilde{C}_{\bar{k}^S+1}$ is sufficiently high so that $\tilde{C}_{\bar{k}^S+1}$ will never bind, and if under the discount factor $\hat{\beta}_1$ the most vulnerable surviving country along the optimal path survives (under conditions of indifference ensured by the choice of L and r) at the brink, then a slightly higher discount factor $\hat{\beta}_2$ will lead to additional countries collapsing along the optimal path.

Finally we note that there are other parameterizations of the model that could also generate this possibility. Suppose, for example, that λ is strictly positive but that the brink point for country $\bar{k}^S + 1$ is close to $\tilde{C}_{\bar{k}^S}$, and in particular suppose that model parameters are such that $\tilde{C}_{\bar{k}^S+1} = \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1)$ and that the brink at $\tilde{C}_{\bar{k}^S+1}$ would not be exceeded along the optimal path. Then, with $\tilde{C}_{\bar{k}^S+1}$ reached at the end of the period in which $\tilde{C}_{\bar{k}^S}$ is exceeded, the utility of each subsequent generation would be a constant $\tilde{V}$ (now this will be so because emissions will be held constant at the level that holds the carbon stock at the level $\tilde{C}_{\bar{k}^S+1}$, and with the carbon stock fixed, λ need not be equal to zero for this result); and if $\tilde{C}_{\bar{k}^S+1}$ is itself close enough to $\tilde{C}_{\bar{k}^S}$ then $\left[B(\bar{c}_{\bar{k}^S}(\hat{\beta}_1)) - \lambda\tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1) + \hat{\beta}_1 V_{\bar{k}^S}(\hat{\beta}_1, \tilde{C}^+_{\bar{k}^S}(\hat{\beta}_1))\right]$ can be brought arbitrarily close to $\frac{1}{1-\hat{\beta}_1}\tilde{V}$, ensuring that (7.19) can be made to hold.

8. References

Barrett, Scott (1994), "Self-enforcing international environmental agreements," *Oxford Economic Papers* 46: 878–894.

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

Barrett, Scott (2013), "Climate treaties and approaching catastrophes," *Journal of Environmental Economics and Management* 66(2): 235-250.

Barrett, Scott and Astrid Dannenberg (2018), "Coercive Trade Agreements for Supplying Global Public Goods," mimeo.

Battaglini, Marco, Salvatore Nunnari and Thomas Palfrey (2014), "Dynamic Free Riding with Irreversible Investments," *The American Economic Review* 104(9): 2858-2871.

Battaglini, Marco and Bård Harstad (2016), "Participation and Duration of Environmental Agreements," *Journal of Political Economy* 124(1): 160-204.

Besley, Timothy and Avinash Dixit (2019), "Environmental catastrophes and mitigation policies in a multiregion world," *Proceedings of the National Academy of Sciences* 166 (12): 5270-5276.

Brander, James and Taylor, M. Scott (1998), "The Simple Economics of Easter Island: A Ricardo-Malthus Model of Renewable Resource Use," *American Economic Review* 88(1): 119-38.

Caplin, Andrew and John Leahy (2004), "The Social Discount Rate," *Journal of Political Economy* 112(6): 1257-1268.

Carraro, C., and D. Siniscalco (1993), "Strategies for the International Protection of the Environment," *Journal of Public Economics* 52(3): 309-28.

Dixit, Avinash (1987), "Strategic Aspects of Trade Policy," in Truman Bewley (ed.), **Advances in Economic Theory: Fifth World Congress**, Cambridge University Press.

Dutta, Prajit K., and Roy Radner (2004), "Self-Enforcing Climate-Change Treaties," *Proceedings of the National Academy of Science* 101: 4746–51.

Dutta, Prajit K., and Rangarajan K. Sundaram (1993), "The Tragedy of the Commons?," *Economic Theory* 3(3): 413-426.

Esty, Daniel C. (1994), **Greening the GATT: Trade, Environment, and the Future**, Institute for International Economics, Washington DC.

Farhi, Emmanuel and Ivan Werning (2007), "Inequality and Social Discounting," *Journal of Political Economy* 115(3): 365-402.

Feng, Tangren and Shaowei Ke (2018), "Social Discounting and Intergenerational Pareto," *Econometrica* 86(5): 1537-1567.

Harstad, Bård (2012), "Climate Contracts: A Game of Emissions, Investments, Negotiations, and Renegotiations," *The Review of Economic Studies* 79(4): 1527–1557.

Harstad, Bård (2021), "Trade and Trees," mimeo, University of Oslo.

Harstad, Bård (forthcoming), "Pledge-and-Review Bargaining: From Kyoto to Paris," *Economic Journal*.

Jenkins, Jesse D. (2014), "Political economy constraints on carbon pricing policies: What are the implications for economic efficiency, environmental efficacy,and climate policy design?," *Energy Policy* 69: 467-477.

John, Andrew and Rowena A. Pecchenino (1997), "International and Intergenerational Environmental Externalities," *Scandinavian Journal of Economics* 99(3): 371–387.

Jones, Aled and Nick King (2021), "An Analysis of the Potential for the Formation of 'Nodes of Persisting Complexity'," *Sustainability* 13, 8161: https://doi.org/10.3390/su13158161.

Kolstad, C. D., and M. Toman (2005), "The Economics of Climate Policy," **Handbook of Environmental Economics** 3: 1562-93.

Kotlikoff, Laurence, Felix Kubler, Andrey Polbin, Jeffrey Sachs and Simon Scheidegger (2021a), "Making Carbon Taxation a Generational Win Win," *International Economic Review* 62(1): 3-46.

Kotlikoff, Laurence, Felix Kubler, Andrey Polbin and Simon Scheidegger (2021b), "Can Today's and Tomorrow's World Uniformly Gain from Carbon Taxation?," mimeo, October.

Lemoine, Derek and Ivan Rudik (2017), "Steering the Climate System: Using Inertia to Lower the Cost of Policy," *American Economic Review*, 107(10): 2947–2957.

Lenton,, Timothy M., Hermann Held, Elmar Kriegler, Jim W. Hall, Wolfgang Lucht, Stefan Rahmstorf and Hans Joachim Schellnhuber (2008), "Tipping elements in the Earth's climate system," *PNAS* 105(6): 1786-1793.

Lenton, Timothy M., Rockstrom, Johan, Gaffney, Owen, Rahmstorf, Stefan, Richardson, Katherine, Steffan, Will and Hans Joachim Schellnhuber (2019), "Climate tipping points – too risky to bet against," *Nature* (Comment) 575, November 28: 592-595.

Lustgarten, Abrahm (2020), "How Climate Migration Will Reshape America: Millions will be displaced. Where will they go?," *The New York Times Magazine*, September 15.

Maggi, Giovanni (2016), "Issue Linkage," in K. Bagwell and R.W. Staiger (eds.), **The Handbook of Commercial Policy**, vol. 1B, Elsevier.

Mattoo, Aaditya and Arvind Subramanian (2013), **Greenprint: A New Approach to Cooperation on Climate Change**, Center for Global Development, Washington, D.C.

Millner, Antony and Geoffrey Heal (2021), "Choosing the Future: Markets, Ethics, and Rapprochement in Social Discounting," NBER Working Paper No 28653, April.

Nordhaus, William D. (2015), "Climate Clubs: Overcoming Free-riding in International Climate Policy," *American Economic Review* 105(4): 1339-70.

Pindyck, Robert S. (2020), "What We Know and Don't Know about Climate Change, and Implications for Policy," NBER Working Paper No 27304.

Rockstrom, Johan et al. (2009), "A safe operating space for humanity," *Nature* 461(24): 472-475.

Russonello, Giovanni (2021), "What's Driving the Surge at the Southern Border?," *The New York Times*, April 5.

Wallace-Wells, David (2019), **The Uninhabitable Earth: Life After Warming**, Tim Duggan Books, New York.

Zaki, Jamil (2019), "Caring about tomorrow: Why haven't we stopped climate change? We're not wired to empathize with our descendants," *The Washington Post* (Outlook), August 22.

Figure 1: ICA, Planner and Noncooperative Outcomes (High $\tilde{C}$)

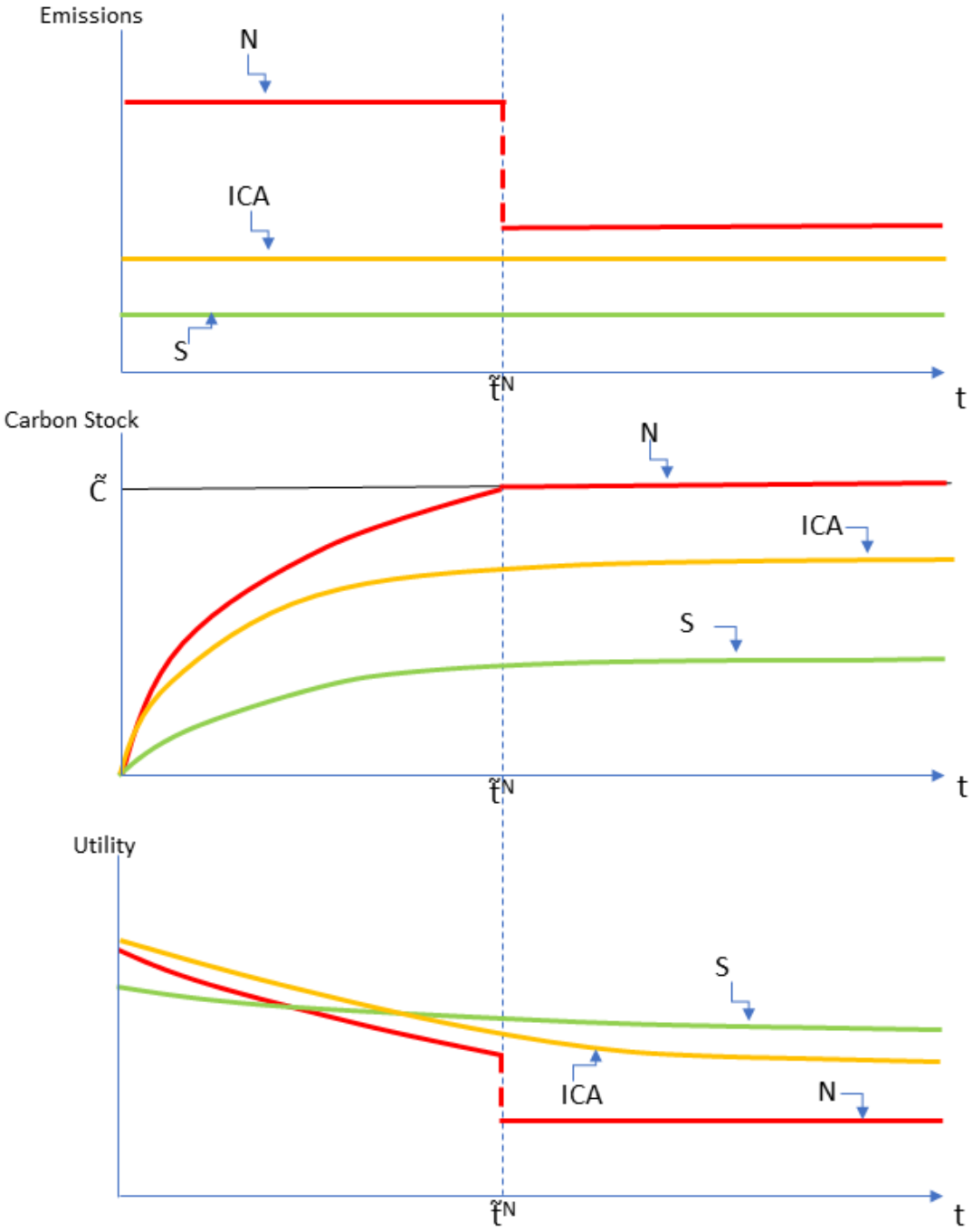

Figure 2: ICA, Planner and Noncooperative Outcomes (Intermediate $\tilde{C}$)

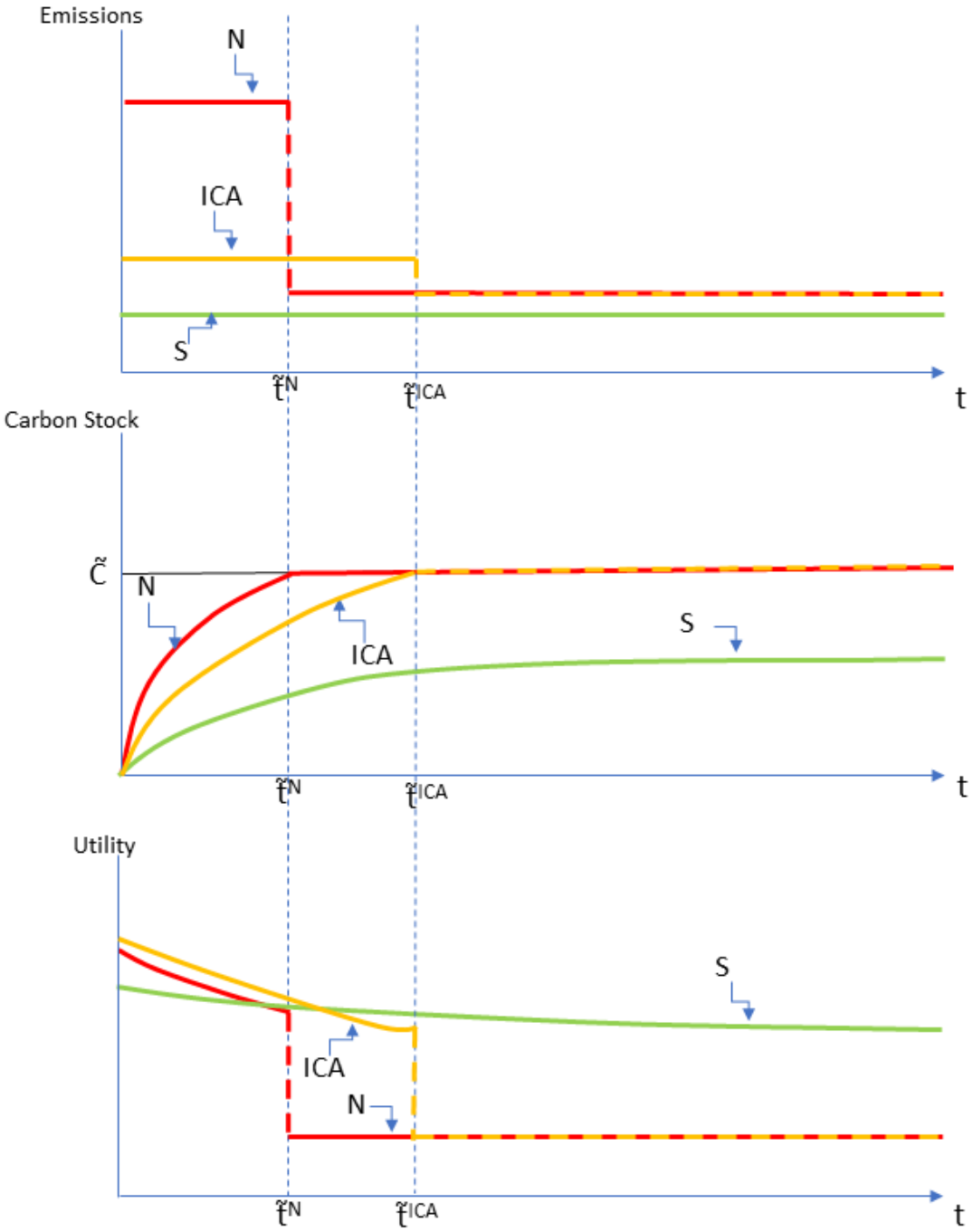

Figure 3: ICA, Planner and Noncooperative Outcomes (Low $\tilde{C}$)

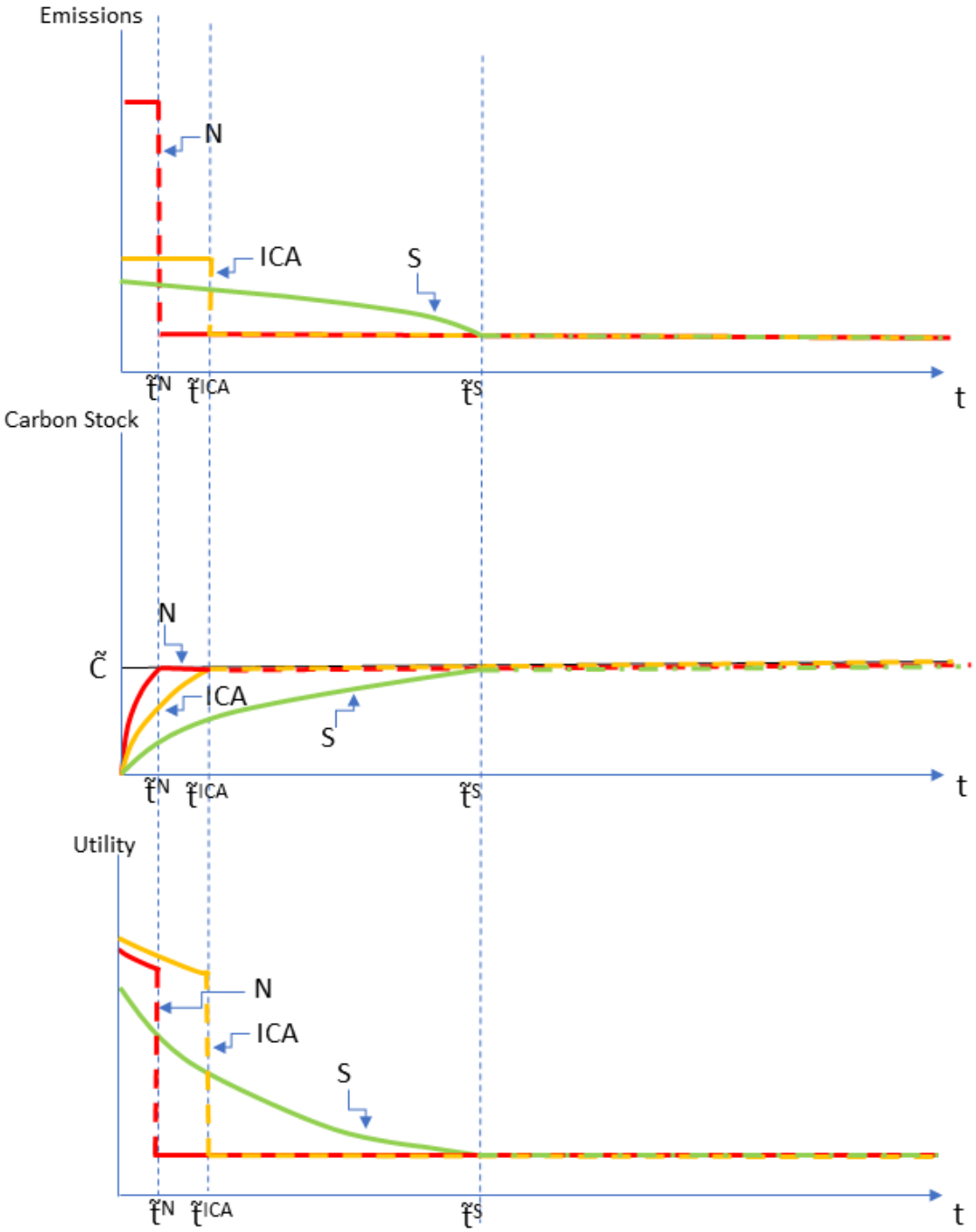

Figure 4(a): Country A's period-1 reaction function for $\beta > 0$ and a fixed $\tilde{C}$ in the two-period Common-Brink Scenario

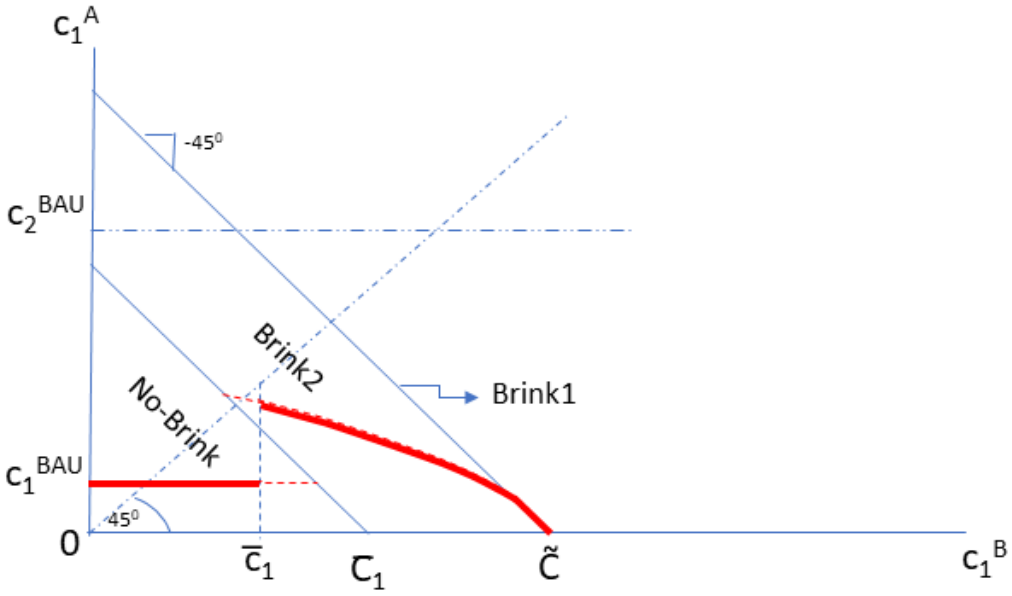

Figure 4(b): Equilibrium period-1 emissions for $\beta > 0$ as a function of $\tilde{C}$ in the two-period Common-Brink Scenario

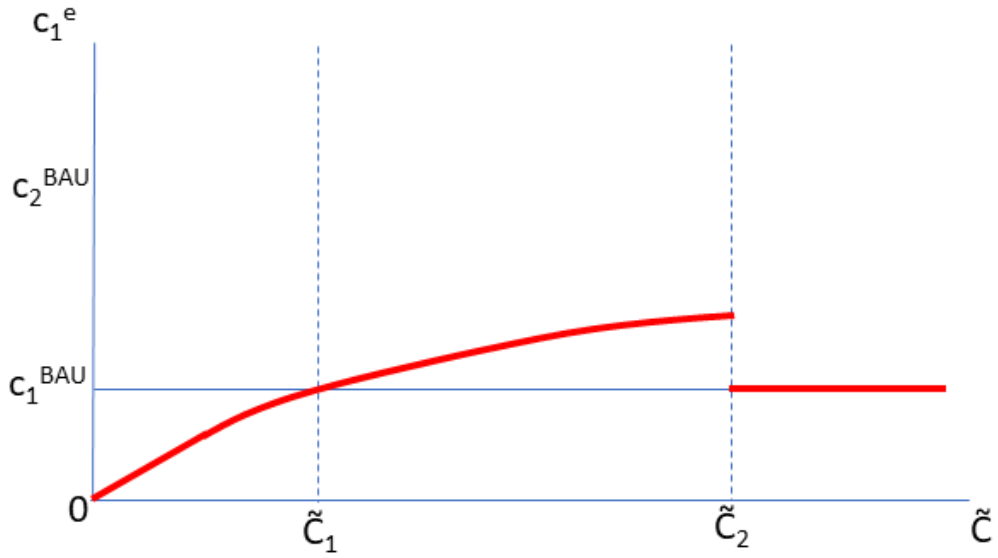

Figure 5(a): Common Brink, Noncooperative Equilibrium

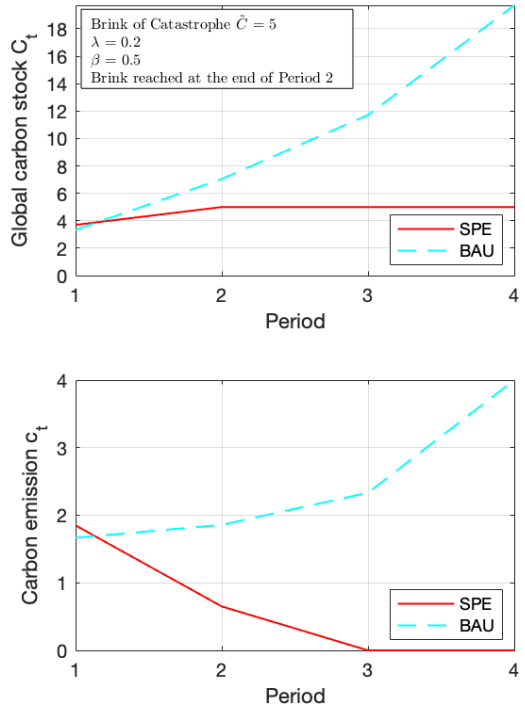

Figure 5(b): Common Brink, Noncooperative Equilibrium

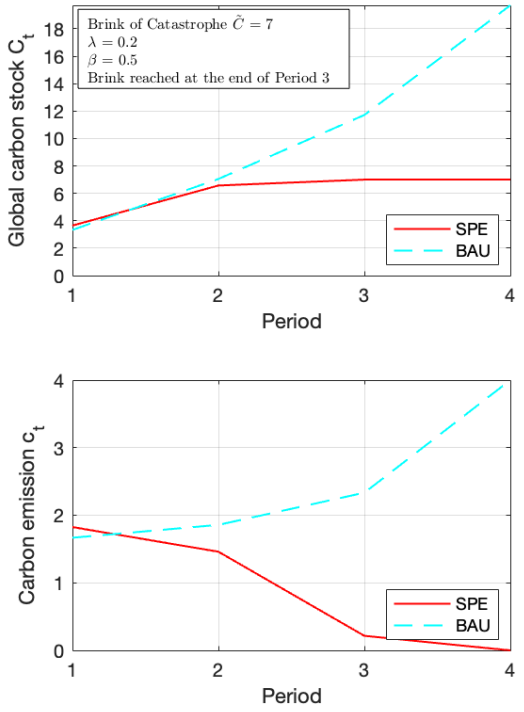

Figure 5(c): Common Brink, Noncooperative Equilibrium

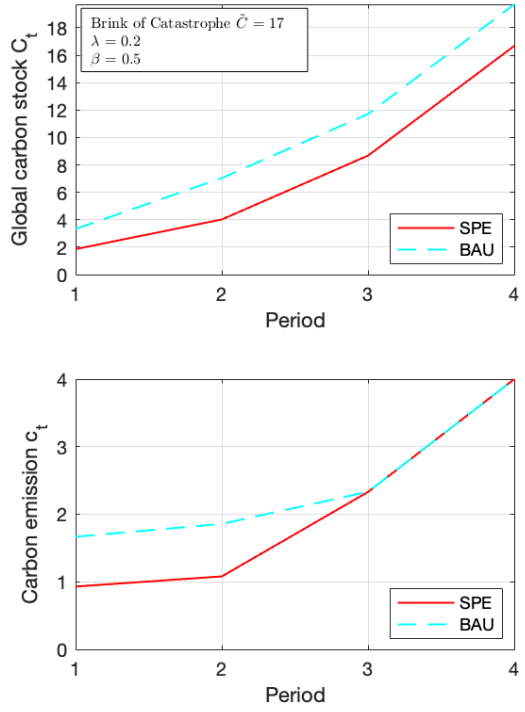

Figure 5(d): Common Brink, Noncooperative Equilibrium