The University of Manchester

Economics

Discussion Paper

Series

EDP-2302

International Cooperation and Kantian Moral Behaviour – Complements or Substitutes?

Alistair Ulph, David Ulph

May 2023

Economics

School of Social Sciences

The University of Manchester

Manchester M13 9PL

International Cooperation and Kantian Moral Behaviour – Complements or Substitutes?

Alistair Ulph
(School of Social Sciences, University of Manchester)
and
David Ulph
(School of Economics & Finance, University of St Andrews)

Abstract

Faced with a global emissions problem such as climate change we know that if countries' emissions decisions are made in an independent and self-interested fashion the outcome can be very far from optimal. One proposed solution is to have countries act more morally by co-operating and so taking account of the impact of their emissions decisions on the welfare of other countries. However, if the decision to co-operate is made in a self-interested fashion the standard non-cooperative model of IEAs yields the *pessimistic conclusion* that the more serious the environmental problem the smaller will be the equilibrium membership of an IEA. Our paper examines the implications for emissions, IEA membership and welfare of assuming that countries make both emissions and IEA membership decisions in the alternative moral fashion of acting as imperfect Kantians as defined by Alger and Weibull (2013). A similar approach has been taken in Eichner and Pethig (2022) who show that the grand coalition (and first-best) can be achieved when countries have a weight on Kantian behaviour greater than a critical value below 2/3. We argue that their approach to modelling the membership decision of imperfect Kantians is problematic and propose an alternative approach. We show that (i) for any weight attached to Kantian behaviour, the equilibrium level of IEA membership and resulting global welfare is higher using our model; (ii) consequently achieving the grand coalition and hence first-best does not require such a high weight on Kantian behaviour; (iii) acting cooperatively and in a Kantian fashion are complementary rather than substitute moral approaches to achieving the first best.

May 2023.

JEL classification: C72, Q50, Q58

Key words: international environmental agreements, moral behaviour, Kantian ethics

Acknowledgements: We are very grateful to Thomas Eichner and Rudiger Pethig for our many discussions while we developed our related, but different, approaches to modelling IEAs with Kantian behaviour. We are also grateful to Michael Finus, Karine Nyborg, and referees of an earlier version of this paper for comments. We are especially grateful to Rahul Prakash and Ji Zhou for research assistance on the numerical calculations, as well as their comments on earlier drafts.

1 Introduction

Climate scientists predict that with high probability there will be potentially catastrophic damage unless global emissions of greenhouse gases are rapidly reduced down to net zero emissions per annum by 2050 (UNEP 2019). Economists emphasise that, in the absence of a global government, the global externality nature of the problem means that individual countries acting independently and in a purely self-interested fashion will make significantly smaller reductions in emissions than those required to achieve the global optimum solution. Acting in a purely self -interested fashion means that (i) a country cares only about its own welfare; (ii) it chooses the action that maximise this welfare without any regard to the decisions of other countries.

A long-standing proposal for addressing this problem has been the creation of an International Environmental Agreement (IEA) whereby countries act in a cooperative fashion and agree to set their individual emissions to achieve the best *collective* outcome for all participating countries. By making emissions decisions in this way countries are acting neither independently nor in the first sense of self-interest, so, in that sense, could be said to be acting more morally with regard to their emissions decisions. Indeed, it is well known that if all countries were to enter such an agreement, then the global optimum could be achieved since countries would be effectively internalising the externality[1].

The problem arises if countries make the decision to join an IEA in a self-interested fashion[2]. For then the workhorse two-stage non-cooperative game model[3], yields the *pessimistic conclusion* that the more serious is the environmental problem, (i.e. the greater is the gap between the non-cooperative and the fully cooperative outcomes) the smaller will be the size of a stable IEA. The intuition is that, for any membership size, fringe countries are always better off than coalition countries, and, the more serious is the environmental problem, the larger are the gains to fringe countries from free-riding on the emissions reductions of IEA members[4,5].

[1] See Finus and Caparros, (2015) for a recent survey of the literature.
[2] By definition the decision to join an IEA has to made independently otherwise there would be some prior agreement.
[3] In this model countries, acting in a self-interested fashion, decide first whether to join or leave a coalition and then their emissions – see, for example, Barrett (1994).
[4] Similar results are derived in Hoel (1992), Carraro and Siniscalco (1993), Rubio and Casino (2002), Finus and Rundshagen (2001), De Cara and Rotillon (2001), and Eichner and Pethig (2015)..
[5] A related strand of literature has assumed that in the Stage 2 emissions game, the IEA acts as Stackelberg leader with respect to the fringe and shows that it may be possible to achieve the grand coalition (Diamantoudi and Sarzetakis 2006, Rubio and Ulph 2006, Nkuyia, 2020, Finus, Furini and Rohrer, 2021a, 2021b. In this paper we assume a Nash equilibrium in the emissions game.

There have been two approaches to obviating this pessimistic conclusion. The first explores richer forms of cooperative behaviour with respect to membership decisions[6].

The second approach continues to employ a non-cooperative Nash equilibrium to study membership decisions, but assumes that countries do not pursue their own self-interest but act in a moral fashion when making both their membership and emissions decisions.[7]

A number of papers have studied how different forms of moral behaviour might allow IEAs to achieve better outcomes - in some cases the grand coalition[8]. In this paper we study non-cooperative IEAs when countries adopt an *imperfect Kantian* form of moral reasoning, in which they give some weight, $\kappa, 0 \leq \kappa \leq 1$, to the Kantian *categorical imperative*[9] (Kant, 1785) to "act only according to that maxim through which you can at the same time will that it become a universal law"[10], which we call acting as a *perfect Kantian,* and the weight $(1-\kappa)$ to acting in a self-interested fashion[11]. As such this form of behaviour is clearly not self-interested in the second sense identified above. With one exception which we mention below, this approach has not so far been studied in the context of IEAs.

To understand how outcomes in terms of emissions and welfare might depend on whether it is the emissions or IEA membership decision to which this calculus is applied we adopt the general approach of allowing countries to apply different Kantians weights to each type of

[6] The γ-core model (see for example, Chander and Tulkens 1997), assumes that when a country leaves a coalition all other countries leave; under appropriate assumptions, this punishment is sufficient to sustain the grand coalition. The literature on farsighted equilibria (see, for example, de Zeeuw, 2008, Diamantoudi and Sartzetakis 2015, 2018) takes an approach intermediate between the fully cooperative and non-cooperative approaches. Farsighted countries ask: if one or more leave/join an IEA of a given membership. will that trigger other countries to modify their membership decisions; if so, will that in turn trigger further changes in membership decisions. An IEA is farsightedly stable if there is no finite chain of membership decisions that would lead to another stable IEA. Under appropriate assumptions this can result in all countries being members of a stable IEA – the grand coalition.

[7] This approach draws on empirical evidence that agents act in a moral fashion when their actions have implications for the well-being of others. For evidence on ethical consumer behaviour see Sudbury-Riley and Kohlbacher (2016) and White, Habib and Hardy (2019). Evidence on how ethical behaviour influences government policymaking is found in Nawrotzki (2012), Kamarack (2019) and Romeijn (2020). The link between governments' willingness to act morally and their electorates willingness is emphasised in Bernauer et al (2016)

[8] Specific forms of moral behaviour affecting IEAs include: modesty (Finus and Maus, 2018); preference for equity (Lange and Vogt, 2003, Vogt, 2016, Rogna and Vogt, 2020); altruism (van der Pol et al,. 2012), reciprocity (Nyborg, 2018b, Bucholz et al, 2018). Much of the evidence for these forms of moral behaviour stem from social surveys or laboratory experiments. See also van Long, 2016, Dasgupta, Southerton, Ulph and Ulph, 2016, and Nyborg 2018a for useful surveys.

[9] This is the 'Universal Law' formulation of Kant's categorical imperative. He proposed two other formulations: the 'Humanity as an End in Itself' and the 'Kingdom of Ends'.

[10] Since one of the choices to which we want to apply this approach is that of emissions levels and since it only makes sense to have a given level of emissions be a universal law if countries are identical, for the purposes of this paper we confine attention to the case of identical countries. We leave it future research to investigate Kantian emissions policies for non-identical countries.

[11] This formulation of imperfect Kantian reasoning draws on the seminal analysis of Alger and Weibull (2013, 2016, 2020) who use evolutionary game theory models of aggregative games with assortative matching to show that the *unique* evolutionary stable preferences are imperfect Kantian.

decision. So κ_ε , $0 \leq \kappa_\varepsilon \leq 1$ (resp. κ_μ , $0 \leq \kappa_\mu \leq 1$) denotes the Kantian weight to the emissions (resp. membership) decision.

We are interested in the question of whether moral behaviour in the form of (imperfect) Kantian behaviour is a substitute or complement for co-operative behaviour. To address this question, we start with the second (emissions) stage of the game and analyse how equilibrium emissions of a coalition country, a fringe country, and an 'average country' vary with respect to the size of the coalition and the Kantian weight on emissions. We show that if countries act as perfect Kantians then, whatever the size of the coalition, the first-best is achieved, while, if there is full cooperation (the grand coalition) then, whatever the Kantian weight on emissions, once again the first-best is achieved. In that sense they are substitute forms of behaviour. More generally we would expect average equilibrium emissions to decline with respect to both factors, and, we show that this is the case, albeit using specific functional forms[12].

Turning to the first (membership) stage of the game we examine what values of the Kantian weights on emissions and membership will enable the grand coalition – and hence first-best to be achieved. We show that this can arise with Kantian weights that are positive but significantly less than one. So, in that sense a Kantian form of behaviour promotes co-operation making the two types of behaviour complements[13].

Our paper relates to the recent paper by Eichner and Pethig (2022). However it differs in a number of respects. First, we explicitly capture the idea that coalition members have as their objective the total welfare of the coalition – and so are not self-interested in the first sense. Second, we employ the imperfect Kantian approach to the membership and emissions decisions in a *nested* fashion which allows each stage to be modelled using the binary imperfect Kantian payoff function derived by Alger and Weibull (2013, 2016, 2020). Thus, at Stage 1 countries make their membership decisions in an imperfect Kantian fashion, recognising that at Stage 2 both coalition and fringe countries will make their emissions decisions in an imperfect Kantian fashion. Third, and relatedly, we use a different approach when modelling the (imperfect Kantian) membership decision. Here, when a coalition (fringe) country assumes that all other countries have made the same membership decision as it has made it also assumes that *the appropriate second stage emissions will apply* (fully-cooperative for a coalition country, non-cooperative for a fringe country). Fourth, since it is difficult to derive results using general functional forms, we have to turn to special functional forms, but ours is more general than theirs since we assume the damage cost function is quadratic, not linear[14]. Fifth while Eisner

[12] However, matters are more complex when we look at the emissions of fringe and coalition countries. \we show that a fringe country's emissions increase with membership, while, for specific values of parameters, a coalition country's emissions also increase with *both* factors.

[13] The poor track record of successive COP meetings in tackling climate change has led some commentators (for example, Nowakowski and Oswald 2020) to argue that environmental economists should focus more on how to change individual behaviour to act more morally, and less on designing complicated theoretical interventions, such as the farsighted model of IEAs or tradable emission permits However, leaving aside the point that co-operation involves some form of moral behaviour, this conclusion suggests that it is wring to pose these as substitute rather than complementary approaches.

[14] Results using the same special case as Eichner and Pethig (2022) are available in an online Appendix.

and Pethig focus solely on how the different values of the Kantian weights affect the equilibrium size of the IEA we look at the impact on the gap in welfare between the non-cooperative outcome and the first best.

,

Finally, given all these differences in approach, we show that for the same values of Kantian weights, (a) our model results in higher equilibrium coalition membership and global welfare; (b) consequently the grand coalition, and hence the first-best outcome, can be achieved with lower Kantian weights.

The structure of the paper is as follows. Section 2 sets out our theoretical analysis. After setting out our model and some key benchmark equilibria, we analyse the non-cooperative equilibrium with Kantian behaviour; the following sub-section summarises the analysis of IEAs with our simple model of Kantian behaviour, establishing the pessimistic conclusion. The final theoretical sub-section sets out two more general models of IEAs with Kantian behaviour; after summarising the results in Eichner and Pethig (2022), we set out our general model and show how it addresses some issues we raise about Eichner and Pethig's model. In Section 3 we analyse the outcomes that arise for the special case of quadratic benefit function and damage cost functions, comparing the results from our general model with those derived from Eichner and Pethig[15]. Section 4 concludes with suggestions for further work.

2 Theoretical Results.

There are $n > 3$ identical countries, $i \in N, N = \{1, \dots, n\}$, where n is a large number[16]. We denote by e_i the level of emissions of country i of a global pollutant, such as greenhouse gases, and $\boldsymbol{e} = (e_1, \dots e_i, \dots e_n)$ the vector of emissions by all countries. The associated level of welfare of country i is:

$$W_i(\boldsymbol{e}) \equiv B(e_i) - D[\textstyle\sum_{j\in N} e_j]. \qquad (1)$$

$B(e_i)$ is the level of net benefit (excluding environmental damages) country i derives from whatever production and consumption decisions give rise to its emissions e_i, and $D(\sum_{j\in N} e_j)$ is the level of damage costs it incurs from global emissions, $\sum_{j\in N} e_j$. We make the standard assumptions that $B(0) \geq 0,\ B'(.) > 0, B''(.) < 0, D(0)0, D'(.) > 0, D''(.) \geq 0.$

We now set out two benchmark outcomes. The *first-best*, or *social optimum*, is achieved when the emissions of each country i , e_i^{SO}, are chosen to maximise total welfare $\sum_{j\in N} W_j(\boldsymbol{e})$, giving rise to f.o.c.:

$$B'\left(e_i^{SO}\right) = \textstyle\sum_{j\in N} D'\left(\sum_{k\in N} e_k^{SO}\right), \quad i \in N \qquad (2a)$$

Imposing symmetry, $e_i^{SO} = e^{SO}$, satisfying:

[15] Eichner and Pethig (2022) produce numerical results using quadratic benefit function and *linear* damage cost function. Our Online Appendix presents numerical results for our model using a linear damage cost function.
[16] In Section 3 we will present numerical results where we assume $n = 100$.

$$B'(e^{SO}) = nD'(ne^{SO}) \quad (2b)$$

The resulting common first-best, or socially-optimal, welfare is:

$$W^{SO} = B(e^{SO}) - D(ne^{SO}). \quad (2c)$$

If countries act in a fully-cooperative fashion, whereby each country chooses the emissions that would maximise total welfare of all countries, it is clear that, in the symmetric model, the fully-cooperative level of emissions for a country, e^{FC}, is identical to the social optimum, $e^{FC} = e^{SO}$, with resulting welfare $W^{FC} = W^{SO}$.

If each country acts in a non-cooperative fashion, choosing its emissions to maximise its own welfare, taking as given the emissions of other countries, the resulting non-cooperative equilibrium emissions, $e_i^{NC}, i \in N$, satisfy:

$$B'\left(e_i^{NC}\right) = D'\left(\sum_j e_j^{NC}\right), \quad i \in N \quad (3a)$$

Imposing symmetry, $e_i^{NC} = e^{NC}, \ i \in N$, yields:

$$B'(e^{NC}) = D'(ne^{NC}) \quad (3b)$$

with resulting non-cooperative welfare:

$$W^{NC} = B(e^{NC}) - D(ne^{NC}) \quad (3c).$$

Given our assumptions, it is clear that $e^{NC} > e^{SO}, \ W^{NC} < W^{SO}$. Since we are interested in challenging global environmental issues such as climate change, we assume that these differences are large.

We now turn to the issue of how far the gaps in emissions and welfare between the non-cooperative equilibrium and the social optimum, i.e $(e^{NC} - e^{SO})$, $(W^{SO} - W^{NC})$ might be closed by countries (a) acting in a more moral (Kantian) fashion; (b) seeking to form an IEA; (c) doing both.

2.1 Non-Cooperative Equilibrium When Countries Act in an Imperfect Kantian Fashion

Following Alger and Weibull (2013, 2016, 2020), we denote by $\kappa, 0 < \kappa \leq 1$ the weight a country attaches to the payoff it would get if it acted as a *perfect Kantian*, and a weight $(1 - \kappa)$ to the payoff it would get if it acted in a self-interested fashion. A *perfect Kantian* country *i*, acting non-cooperatively, asks what emissions it should set, *acting on the Kantian hypothesis* that all other countries choose the same level of emissions as it. Thus, the *behaviour* of a perfect Kantian country *i* can be characterised by saying it acts to maximise its *perfect Kantian payoff* function:

$$\Pi_i^{NCK}(\boldsymbol{e}; 1) = B(e_i) - D\left(\sum_{j \in N} e_i\right) = B(e_i) - D(ne_i). \quad (4)$$

We emphasise that while the *perfect Kantian payoff function* is used to characterise the *behaviour* of countries acting in a Kantian fashion, we will continue to assess the well-being of countries using their welfare functions defined in (1).

However, just as it is unrealistic that independent countries would act in a fully cooperative fashion, it is also unrealistic to assume that countries act as perfect Kantians. Therefore, we explore the implications of having countries act non-cooperatively, but as *imperfect Kantians.* Following Alger and Weibull (2013, 2016, 2020), an *imperfect Kantian* country gives a *Kantian weight,* κ, $0 \leq \kappa \leq 1$, to acting in a perfect Kantian fashion, and weight $(1-\kappa)$ to acting in a self-interested non-cooperative fashion. An imperfect Kantian country *i* seeks to maximise its *imperfect Kantian payoff function:*

$$\Pi_i^{NCK}(\boldsymbol{e};\kappa) = \kappa\left[B(e_i) - D\left(\sum_{j\in N} e_i\right)\right] + (1-\kappa)\{B(e_i) - D\left[e_i + \sum_{j\in N, j\neq i} e_j\right]\}$$

$$= B(e_i) - \kappa D(ne_i) - (1-\kappa)D[e_i + \sum_{j\in N, j\neq i} e_j] \qquad (5a)$$

We denote the resulting equilibrium emissions by $e_i^*(\kappa)$ which satisfies:

$$B'[e_i^*(\kappa)] = \kappa n D'[ne_i^*(\kappa)] + (1-\kappa)D'[e_i^*(\kappa) + \sum_{j\in N, j\neq i} e_j^*(\kappa)] \quad i = 1, \ldots, n \quad (5b)$$

Imposing symmetry, the non-cooperative Kantian equilibrium level of emissions, which we denote $e^*(\kappa)$, solves

$$B'[e^*(\kappa)] = [\kappa n + (1-\kappa)]D'[ne^*(\kappa)] \qquad (5c)$$

with the resulting *imperfect Kantian non-cooperative* equilibrium payoff and welfare, which we denote by $\Pi^*(\kappa), W^*(\kappa)$ respectively:

$$\Pi^*(\kappa) = W^*(\kappa) = B[e^*(\kappa)] - D[ne^*(\kappa)] \qquad (5d)$$

It is clear from (5c) that a country acting as a perfect Kantian would choose emissions $e^*(1) = e^{SO}$, while a country acting as a non-Kantian would choose emissions $e^*(0) = e^{NC}$. So, the first best can be achieved *either* by having all countries act in a fully cooperative fashion in setting emissions, *or* by having all countries acting non-cooperatively but as perfect Kantians. If countries are perfect Kantians, there is no need to try to form an IEA: seeking to get countries to act morally, as perfect Kantians, is a *substitute* to seeking to get all countries to join an IEA as an approach for tackling climate change.

Result 1.

In the imperfect Kantian non-cooperative equilibrium, as κ *increases from 0 to 1*,

- *(i)* $e^*(\kappa)$ *falls from the conventional non-Kantian non-cooperative level of emissions* $(e^*(0) = e^{NC})$ *to the first-best level of emissions* $(e^*(1) = e^{FC} = e^{SO})$*;*
- *(ii) the imperfect Kantian payoff and welfare,* $\Pi^*(\kappa) = W^*(\kappa)$*, increase from the non-cooperative level* $(\Pi^*(0) = W^*(0) = W^{NC})$ *to the first-best level* $(\Pi^*(1) = W^*(1) = W^{SO})$.

2.2 IEAs with Imperfect Kantian Behaviour

We employ the two-stage non-cooperative model of IEAs, stemming from Carraro and Siniscalco (1993), Barrett (1994), in which in Stage 1 countries determine the equilibrium membership of a coalition or fringe and in Stage 2 they determine their equilibrium emissions. We denote by *C* (*F*), the set of countries which belong to the coalition (fringe) respectively

following in Stage 2, where $C \cup F = N$, and by C_{-i} (F_{-i}) the set of countries belonging to the coalition (fringe) excluding country i. For the case of identical countries, we denote by m the number of countries in the coalition, so, n - m is the number of countries in the fringe; if m = n, all countries are in the coalition, while if m = 1 all countries are in the fringe. In this sub-section we examine four models, which depend on whether or not countries act as imperfect Kantians with respect to emissions, membership, none, or both. We denote by $\kappa, 0 \leq \kappa \leq 1$ the Kantian weight a country acting as an imperfect Kantian attaches to acting as a perfect Kantian with respect to emissions, membership or both. As noted above, the first three models appear also in Eichner and Pethig (2022).

2.2.1 IEAs: Countries Are Self-Interested (Non-Kantian $\kappa = 0$) – Model *v*

The payoff function for a typical country *i* is given by[17]:

$$\Pi_i^v(\boldsymbol{e}; 0) = B(e_i) - D(\sum_{j\in C} e_j + \sum_{k\in F} e_k) \qquad i \in N,$$

where the term 0 in $\Pi_i^v(\boldsymbol{e}; 0)$ denotes the value of κ.

Stage 2: Equilibrium Emissions

If country *i* is a member of some given coalition, *C*, it chooses its emission e_i to maximise $B(e_i) + \sum_{j\in C_{-i}} B(e_j) - \sum_{j\in C} D(\sum_{j\in C} e_j + \sum_{k\in F} e_k)$ with resulting first-order condition:

$$B'(e_i) = \sum_{j\in C} D'(\sum_{j\in C} e_j + \sum_{k\in F} e_k) \quad i \in C \qquad (6a)$$

If country *i* is a member of the fringe, *F*, it chooses its emissions e_i to maximise $B(e_i) - D(\sum_{j\in C} e_j + \sum_{k\in F} e_k)$ with resulting first-order condition:

$$B'(e_i) = D'(\sum_{j\in C} e_j + \sum_{k\in F} e_k) \qquad i \in F \qquad (6b)$$

For the case of identical countries, the first-order conditions become:

$$B'[e^c] = mD'[me^c + (n-m)e^f] \quad i \in C \qquad (6c)$$

$$B'[e^f] = D'[me^c + (n-m)e^f] \quad i \in F \qquad (6d)$$

Solving (6c) and (6d) yields equilibrium emissions $\hat{e}^c(m; 0), \hat{e}^f(m; 0)$; the resulting Stage 2 equilibrium payoffs are:

$$\widehat{\Pi}^j(m; 0) = B\left[\hat{e}^j(m; 0)\right] - D[m\hat{e}^c(m; 0) + (n-m)\hat{e}^f(m; 0)], j = c, f \qquad (6e)$$

From (6a) and (6b), it is clear that, for all m = 2,..,n – 1, $\hat{e}^f(m; 0) > \hat{e}^c(m; 0)$, and hence, from (6e), $\widehat{\Pi}^f(m; 0) > \widehat{\Pi}^c(m; 0)$, the standard free-rider advantage for fringe countries. If m = n, the grand coalition, all countries are coalition members and only (6a) applies, so $\hat{e}^c(n; 0) = e^{SO}, \widehat{\Pi}^c(n; 0) = W^{SO}$. If m = 1, all countries are in the fringe and only (6b) applies, so $\hat{e}^f(1; 0) = e^{NC}, \widehat{\Pi}^f(1; 0) = W^{NC}$.

[17] For the non-Kantian model, the payoff function is equal to the welfare function.

Stage 1: Equilibrium Membership

Equilibrium membership of the IEA, $\hat{m}$, is determined as a non-cooperative equilibrium in which no coalition country would wish to unilaterally leave the current coalition of size $\hat{m}$ and join the fringe (Internal Stability), and no fringe country would wish to unilaterally leave the fringe of size $n - \hat{m}$ and join the coalition (External Stability), i.e.

$$\hat{\Pi}^c[\hat{m}] \geq \hat{\Pi}^f[\hat{m} - 1] \qquad (7a)$$

$$\hat{\Pi}^f[\hat{m}] \geq \hat{\Pi}^c[\hat{m} + 1] \qquad (7b)$$

Equivalently, we define the stability function: $\sigma(m) \equiv \hat{\Pi}^c(m) - \hat{\Pi}^f(m-1)$, and say that a coalition of size $\hat{m}$ is stable iff

$$\sigma[\hat{m}] \geq 0, \;\; \sigma[\hat{m}+1] \leq 0 \qquad (7c)$$

The overall equilibrium for model *v*, where countries are non-Kantians, is characterised by the key outcomes: the size of the stable IEA, $\hat{m}_v$; the resulting equilibrium emissions and *welfare* of coalition and fringe countries: $\hat{e}_v^j = \hat{e}^j[\hat{m}_v; 0], \hat{W}_v^j = \hat{\Pi}^j[\hat{m}_v; 0], j = c, f$.

As we noted in Section 1, it is well known that, for a wide class of functional forms for benefit and damage cost functions, when countries act in a self-interested manner, we get the pessimistic conclusion that, because of the free-rider benefits accruing to fringe countries, the size of the stable IEA, $\hat{m}_v$, is small[18]. Applying this conclusion to problems like climate change involving a large number of countries, may seem to support the argument of Nowakowski and Oswald (2020) environmental economists' focus on how to secure an IEA to tackle climate change, rather than on changing individuals' behaviour, may be misplaced.

2.2.2 IEAs: Countries Are Kantians with Respect to Emissions Only: Model ε

We begin with the case where all countries are *perfect* Kantians with respect to emissions. A typical country *i* has payoff function:

$$\Pi_i^{\varepsilon}(\boldsymbol{e}; 1) = B(e_i) - D(ne_i) \qquad i \in N \qquad (8a)$$

where the term 1 in $\Pi_i^{\varepsilon}(\boldsymbol{e}; 1)$ denotes the value of κ. At Stage 2, whatever coalition, *C*, has formed at Stage 1, a coalition country *i* chooses its emissions, e_i, to maximise:

$$B(e_i) - D(ne_i) + \sum_{j \in C_{-i}}[B(e_j) - D(ne_j)] \;\; i \in C \qquad (8b)$$

The objective function takes this form because, while country *i* seeks to maximise the total payoff of all coalition countries, it has no agency over the emissions decisions of other coalition members. Similarly, a fringe country, *i*, chooses its emissions e_i to maximise:

$$B(e_i) - D(ne_i) \qquad i \in F \qquad (8c)$$

[18] Typically, in the range 2-4, see Diamantoudi and Sartzetakis (2006)

Therefore, at Stage 2, a country which is a perfect Kantian with respect only to emissions will set $\hat{e}^c(m;1) = \hat{e}^f(m;1) = e^{SO}$, the socially optimal level of emissions, irrespective of whether it is a member of the coalition or fringe and independent of the size of the coalition.

It follows that, at Stage 1, the stability function is: $\sigma(m) = 0\ ,\ m = 2, \ldots, n$. Therefore, any m = 2, …,n can be a stable IEA, and we apply equilibrium selection to say that the stable IEA is the grand coalition. Thus, the equilibrium for Model ε when κ = 1 is characterised by $\widehat{m}_\varepsilon = n;\ \hat{e}_\varepsilon^j(1) = \hat{e}^j(n;1) = e^{SO};\ \widehat{W}_\varepsilon^j(1) = \widehat{W}^j(n;1) = W^{SO}; j = c, f$. Thus, with perfect Kantian behaviour with respect to emissions only, the outcome is the same as in Section 2.1 when countries act non-cooperatively; forming an IEA is essentially irrelevant.

In the case where countries are *imperfect* Kantians with respect to emissions only, $0 < \kappa < 1$, country *i*'s payoff function is:

$$\Pi_i^\varepsilon(\boldsymbol{e};\kappa) = B(e_i) - \kappa D(ne_i) - (1-\kappa)D[\textstyle\sum_{j\in C} e_j + \sum_{j\in F} e_j] \quad i \in N \qquad (9a)$$

Eichner and Pethig (2022) noted that (9a) can be written as:

$$\Pi_i^\varepsilon(\boldsymbol{e};\kappa) = \breve{B}(e_i) - \breve{D}[\textstyle\sum_{j\in C} e_j + \sum_{j\in F} e_j] \quad i \in N \qquad (9b)$$

where $\breve{B}(e_i) = B(e_i) - \kappa D(ne_i);\ \ \breve{D}[.] = (1-\kappa)D[.]$. It is clear that $\breve{B}(.), \breve{D}(.)$ have the same properties as *B*(.), *D*[.]. The IEA game where countries are *imperfect* Kantians with respect to emissions only is *isomorphic* to the standard game in 2.2.1 above, and so, again, we get the pessimistic conclusion that the stable IEA with be small. Thus, as Eichner and Pethig (2022) noted, equilibrium membership in the emissions only special case has a knife-edge property. While Kantian behaviour with respect to emissions only leads to the pessimistic conclusion with respect to membership, equilibrium Stage 2 emissions, $\hat{e}^c(m,\kappa), \hat{e}^f(m,\kappa)$ obviously depend on the Kantian weight κ. In Section 4 we will analyse how equilibrium emissions vary with parameters *m a*nd κ; as we would expect, as κ increases, total emissions of all countries will fall and total welfare of all countries will increase, reaching social optimum when $\kappa = 1$, but the story is richer for individual countries. Thus, the implications of this model of IEAs for welfare are significant if κ is large. However, the important point is that, with a small equilibrium membership, the outcome will be only slightly better than in the corresponding non-cooperative Kantian equilibrium.

2.2.3 IEAs: Countries Are Kantians with Respect to Membership Only: Model μ

We start with perfect Kantian behaviour with respect to membership. No matter what coalition, *C*, has formed at Stage 1, at Stage 2 a coalition (fringe) country *i* which is a perfect Kantian with respect to membership only, acts on the counter-factual hypothesis that all other countries have joined the coalition (fringe). Hence, it's payoff function is:

$$\Pi_i^\mu(\boldsymbol{e};1) = B(e_i) - D[\textstyle\sum_{j\in N}(e_j)] \qquad i \in N \qquad (10a)$$

Thus, a perfect Kantian coalition country *i* chooses e_i to maximise:

$$\textstyle\sum_{j\in N} \Pi_i^\mu\ (\boldsymbol{e};1) = B(e_i) + \sum_{j\in N-i} B(e_j) - \sum_{j\in N} D[e_i + \sum_{k\in N_{-i}} e_k]\ \ i \in C \qquad (10b$$

with first-order condition:

$$B'(e_i) = \sum_{j\in N} D'(\sum_{k\in N} e_k) \qquad (10c)$$

With identical countries, each coalition member sets equilibrium emissions $\hat{e}^c(m;1)$ to satisfy

$$B'[\,\hat{e}^c(m;1)] = nD'[n\,\hat{e}^c(m;1)] \Longrightarrow\ \hat{e}^c(m;1) = e^{SO}$$

Therefore, whatever the number of countries in the coalition, $m = 2,\ldots,n$, each coalition country sets the socially optimal level of emission, and believes it will receive Stage 2 equilibrium payoff $\hat{\Pi}^c(m;1) = W^{SO}$.

By a similar argument, at Stage 2, a fringe country *i* which is a perfect Kantian with respect to membership only, acts on the hypothesis that all other countries have joined the fringe, and chooses its emissions to maximise the payoff function:

$$\Pi_i^{\mu}(\boldsymbol{e};1) = B(e_i) - D[e_i + \sum_{j\in N, j\neq i} e_j] \quad i \in F \qquad (10d)$$

Therefore, whatever the number of countries in the fringe, n-m = 1, …, n-1, each fringe country sets the non-cooperative level of emissions: $\hat{e}^f(m;1) = e^{NC}$ and believes it will receive Stage 2 equilibrium payoff $\hat{\Pi}^f(m;1) = W^{NC}$.

For any membership *m*, the Stage 1 stability function is $\sigma(m;1) = W^{SO} - W^{NC} > 0$ .Hence the unique stable coalition is the grand coalition. Put simply, if a country believes all other countries will make the same membership decision as it, then it calculates that it would get a higher payoff by joining the coalition and getting the socially-optimal payoff rather than joining the fringe and getting the non-cooperative payoff.

We now turn to imperfect Kantian behaviour with respect to membership, where countries attach weight κ to acting as a perfect Kantian with respect to membership. Whatever coalition *C* has formed at Stage 1, the payoff function of a coalition (fringe) country *i* is:

$$\Pi_i^{\mu}(\boldsymbol{e};\kappa) = \kappa[B(e_i) - D(\sum_{j\in N}(e_j)] + (1-\kappa)\{B(e_i) - D[\sum_{j\in C} e_j + \sum_{k\in F} e_f]\} \qquad (11a)$$

At Stage 2, a coalition country *i* chooses its emissions e_i to maximise:

$$\kappa[B(e_i) - \sum_{j\in N} D(\sum_{k\in N} e_k)] + (1-\kappa)[B(e_i) - \sum_{j\in C} D(\sum_{j\in C} e_j + \sum_{k\in F} e_f)] \qquad (11b)$$

With identical countries, the first-order condition for a coalition country can be written as:

$$B'(e^c) = \kappa n D'(ne^c) + (1-\kappa)mD'[me^c + (n-m)e^f] \qquad (11c)$$

Similarly, a fringe country *i* chooses its emissions e_i to maximise:

$$\kappa[B(e_i) - D(\sum_{k\in N} e_k)] + (1-\kappa)[B(e_i) - D(\sum_{j\in C} e_j + \sum_{k\in F} e_k)] \qquad (11d)$$

and , with identical countries, the first-order condition can be written as:

$$B'(e^f) = \kappa D'(ne^f) + (1-\kappa)D'[me^c + (n-m)e^f] \qquad (11e)$$

Solving (11c) and (11e) simultaneously yields equilibrium emissions: $\hat{e}^j(m;\kappa)$ *j=c,f.*

It is clear that the first-order conditions (11c) and (11e) for Model μ are not iso-morphic to (6c) and (6d) for the self-interested Model *v,* so we cannot conclude that Model μ yields the

pessimistic conclusion with respect to equilibrium membership. The full analysis of Model μ can be found in Eichner and Pethig (2022). For the special case of quadratic benefit function and linear damage cost function, they find that for Model μ the size of the stable IEA increases as κ increases, and, the grand coalition can be stable with a value of κ which tends to 0.5 as n tends to infinity.

2.2.4 IEAs: Countries Are Kantians with Respect to Emissions and Membership: Model $\varepsilon\mu$

Eichner and Pethig (2022) do not consider this case. A coalition (fringe) country i which is a perfect Kantian with respect to both emissions and membership acts on the hypotheses that all other countries have joined the coalition (fringe) and that they will set the same emissions as it. From (8a), (10a) and (10d) it's Stage 2 payoff function is:

$$\Pi_i^{\varepsilon\mu}(\boldsymbol{e};\ 1) = B(e_i) - D(ne_i)\ \ i \in N. \qquad (12a)$$

A coalition country chooses its emissions, e_i, to maximise the joint payoff of all members of it hypotheses to belong to the coalition, i.e. all countries in N, namely:

$$B(e_i) - D(ne_i) + \sum_{j\in N_{-i}}[B(e_j) - D(ne_j)] \qquad (12b)$$

A fringe country i takes as given the emissions of all other members it hypotheses belong to the fringe, i.e all other countries in N, and chooses its emissions, e_i, to maximise:

$$B(e_i) - D(ne_i) \qquad (12c)$$

From (12b) and (12c), it is clear that, at Stage 2, for whatever coalition (fringe) has actually formed at Stage 1, any country, i, whether a member of the coalition or fringe, will set socially optimal emissions, i.e. $\hat{e}_{\varepsilon\mu}^{c} = \hat{e}_{\varepsilon\mu}^{f} = e^{SO}$. Hence, as in 2.2.2, at Stage 1, any IEA with membership $m = 2,\ldots, n$ is stable and we choose the grand coalition. Thus, the full equilibrium when countries are perfect Kantians with respect to both emissions and membership is the social optimum.

From (9a), (11a) and (11d), a country i which is an imperfect Kantian ($0 < \kappa < 1$) with respect to emissions and membership has the payoff function:

$$\Pi_i^{\varepsilon\mu}(\boldsymbol{e};\ \kappa) = B(e_i) - \kappa D(ne_i) - (1-\kappa)D[\textstyle\sum_{j\in C} e_j + \sum_{j\in F} e_j]\ \ \forall i \in N \qquad (12d).$$

This has the same isomorphism property as (9a), so:

$$\Pi_i^{\varepsilon\mu}(\boldsymbol{e};\kappa\ ) = \breve{B}(e_i) - \breve{D}[\textstyle\sum_{j\in C} e_j + \sum_{j\in F} e_j] \qquad (12e)$$

where $\breve{B}(e_i) = B(e_i) - \kappa D(ne_i)$; $\breve{D}[.] = (1-\kappa)D[.]$. Thus, when countries are imperfect Kantians with respect to both emissions and membership, we again get the pessimistic conclusion with respect to membership, although, as we noted at the end of Section 2.2.2, as the Kantian weight $\kappa \to 1$, emissions and hence welfare tend to the social optimum.

We summarise the outcomes of the models studied in Section 2 as follows:

Result 2

(i) *If countries are non-Kantians with respect to both emissions and membership, for a wide class of functional forms, the size of the equilibrium membership is small and welfare is close to the level in the non-cooperative non-Kantian equilibrium.*

(ii) *If countries act as perfect Kantians with respect to emissions only or with respect to emissions and membership, any coalition with membership m, $2 \leq m \leq n$, is stable and achieves the socially optimal outcome; we select the grand coalition as the stable equilibrium.*

(iii) *If countries act as imperfect Kantians with respect to emissions only, or with respect to emissions and membership, for a wide class of functional forms, equilibrium membership is small, though as the Kantian weight tends to* 1, *emissions and welfare tend to the social optimum.*

(iv) *If countries act as perfect Kantians with respect to membership only, the grand coalition is the unique stable IEA, attaining the socially optimal outcome.*

(v) *If countries act as imperfect Kantians with respect to membership only, the iso-morphism property does not apply. For the special case of quadratic benefits and linear damage costs, as the Kantian weight κ increases, the size of the stable IEA increases, reaching the grand coalition for values of κ which exceeds a critical value which tends to* 0.5 *as n tends to infinity.*

This is a disappointing result. We believe: (a) it is unrealistic to expect that countries would act as Kantians with respect to only emissions or only membership, so we should assume they act as Kantians in making both decisions; (b) it is unrealistic to expect that they act as perfect Kantians, for which the issue of forming an IEA is effectively irrelevant. Hence, the outcome for membership of this simple approach to IEAs with Kantian behaviour is the same pessimistic conclusion as in models of self-interested behaviour. In the next section we consider two approaches which assume that the model of Kantian behaviour with respect to both emissions and membership is richer than that presented above.

2.3 IEAs with Kantian Behaviour – Richer Models.

In this section, it will be useful to distinguish the Kantian weight for decisions on emissions, κ_ε, and the Kantian weight for decisions taken on membership, κ_μ. We recognise that, there is a strong argument that an agent should take the same moral stance to all decisions, implying $\kappa_\varepsilon = \kappa_\mu$; we do not preclude this possibility. One reason for thinking that the weights might differ is that membership and emission decisions involve somewhat different agents: decisions about membership of international partnerships are clearly the remit of national governments; while national governments' policies can strongly influence domestic greenhouse gas emissions in the production, retail and domestic sectors, households' carbon footprints also depend on individual and household decisions such as diet and transport which are less amenable to national government interventions. Thus, we allow for the possibility that $\kappa_\varepsilon \neq \kappa_\mu$, with either $\kappa_\varepsilon < \kappa_\mu$, or $\kappa_\varepsilon > \kappa_\mu$[19].

[19] For example, households' decisions on diet, where to live and where to work, how long to use appliances.

2.3.1 The Model of Eichner and Pethig (2022)

We begin by briefly outlining the approach of Eichner and Pethig (2022). They depart from the Alger and Weibull (2013, 2016, 2020) approach to imperfect Kantian behaviour, which we applied in 2.1. and 2.2 above, whereby the payoff function of country *i* is a weighted average of the perfect Kantian payoff and the non-Kantian payoff. Instead, they use a payoff function which is a weighted average of the payoffs of a perfect Kantian with respect to emissions only, a perfect Kantian with respect to membership only, and a non-Kantian. Thus, a country *i*, whether a member of the coalition or fringe, has the payoff function:

$$\Pi_i^{EP}\left(\boldsymbol{e};\ \alpha, \kappa_\varepsilon, \kappa_\mu\right) = \alpha\Pi_i^{\mu}(\boldsymbol{e}; \kappa_\varepsilon) + (1-\alpha)\Pi_i^{\varepsilon}(\boldsymbol{e}; \kappa_\mu) \qquad (13a)$$

where $0 \leq\ \alpha, \kappa_\varepsilon, \kappa_\mu \leq 1;$ a special parameter subset is (0.5, *κ, κ*). (13a) can be expanded as follows:

i.e. $$\Pi_i^{EP}\left(\boldsymbol{e};\ \alpha, \kappa_\varepsilon, \kappa_\mu\right) = \alpha\kappa_\mu[B(e_i) - D(\textstyle\sum_{j\in N} e_j)] + (1-\alpha)\kappa_\varepsilon[B(e_i) - D(ne_i)]$$

$$+\left[\alpha\left(1-\kappa_\mu\right) + (1-\alpha)(1-\kappa_\varepsilon)\right][B(e_i) - D(\textstyle\sum_{j\in C} e_j + \sum_{k\in F} e_k)] \quad (13b)$$

After studying the special cases of non-Kantian ($\kappa_\varepsilon = \kappa_\mu = 0$ as in Section 2.2.1 above), emissions only ($0 < \kappa_\varepsilon \leq 1, \kappa_\mu = 0$, as in Section 2.2.2 above), and membership only ($0 < \kappa_\mu \leq 1, \kappa_\varepsilon = 0$, as in Section 2.2.3 above), they study their general model, and show, importantly, that the iso-morphism result does not apply. For the special case of quadratic benefit and linear damage cost functions for different sets of parameter values, they show: first, that equilibrium membership, $\hat{m}^{EP}\left(\alpha, \kappa_\varepsilon, \kappa_\mu\right)$, is increasing in all three parameters, and, for the special parameter set, is increasing in *κ*; second, they study parameter values for which the grand coalition can be achieved, and show that, there are critical values of α and κ_μ, denoted $\hat{\alpha}(\kappa_\varepsilon, \kappa_\mu, n)$ $\hat{\kappa}_\mu(\alpha, \kappa_\varepsilon, n)$ respectively, lying between 0 and 1, such that the grand coalition can be achieved if $\alpha \geq \hat{\alpha}(.)$, or $\kappa_\mu \geq \hat{\kappa}_\mu(.)$; for the special parameter set (0.5, *κ, κ*) there is a critical value $\hat{\kappa}(n)$ such that the grand coalition can be achieve if $\kappa \geq \hat{\kappa}(n)$, where $\hat{\kappa}(n)$ tends to 2/3 as *n* tends to infinity.[20]

In the next section we set out our general model, and argue why we believe it provides a more appropriate approach to capturing Kantian behaviour in IEAs. Furthermore, we believe the parameter α to be superfluous. The key ethical parameters are the Kantian weights, $\kappa_\varepsilon, \kappa_\mu$. As we noted in the opening paragraph of Section 2.3, arguments can be advanced for why these might differ from each other. However, it is not clear what *moral* arguments can be advanced for a choice of α other than 0.5, which makes α superfluous. For example, if one accepts arguments for having $\kappa_\mu > \kappa_\varepsilon$, what additional arguments could be advanced for choosing α such that $\alpha\kappa_\mu < (1-\alpha)\kappa_\varepsilon$? We will come back to this issue in Section 3 when we seek to compare the results from our model with those from Eichner and Pethig (2022).

[20]For later purposes, we note that $\hat{\kappa}(100) = 0.6644$.

2.3.2 Our General Model

We now turn to the main section of this paper. Our general model differs from Eichner and Pethig (2022) in two key respects. First, we employ the imperfect Kantian approach to the membership and emissions decisions in a *nested* fashion which allows each stage to be modelled using the binary imperfect Kantian payoff function derived by Alger and Weibull (2013, 2016, 2020). By this we mean that, at Stage 1, countries make their membership decisions in an imperfect Kantian fashion with Kantian weight κ_μ, recognising that at Stage 2 they will make their emissions decisions in an imperfect Kantian fashion with Kantian weight κ_ε. This has the important implication that, unlike Eichner and Pethig (2022), the weights $\kappa_\varepsilon, \kappa_\mu$ do not have to sum to 1. The second important difference from Eichner and Pethig (2022) is that we use a different approach to modelling membership decisions in which a coalition (fringe) country acting as perfect Kantians, assumes that all other countries have made the same membership decision *and chooses the appropriate emissions*. By appropriate we mean that they take account of what they hypothesise about membership and the Kantian weight on emissions.

We now set out the payoff functions for coalition and fringe countries. It follows from the previous paragraph that it is necessary to distinguish between two sets of emission decisions, reflecting the difference between acting as imperfect Kantians with respect to emissions and with respect to membership. We denote by e_i^ε the emissions of country *i* in the emissions game, and by e_i^μ the emissions country *i* would set in the membership game when it assumes that, hypothetically, all other countries have made the same membership decision. Although we distinguish between two sets of emissions, we note shortly that, in *equilibrium*, countries emit only one level of emissions which depends on the equilibrium membership. The payoff function is:

$$\Pi_i\left(\boldsymbol{e}^\varepsilon, \boldsymbol{e}^\mu; \kappa_\varepsilon, \kappa_\mu\right) = \kappa_\mu\{\kappa_\varepsilon[B(e_i^\mu) - D(\textstyle\sum_{j\in N} e_i^\mu)] + (1-\kappa_\varepsilon)[B\left(e_i^\mu\right) - D(\textstyle\sum_{j\in N} e_j^\mu)]\}$$
$$+\left(1-\kappa_\mu\right)\{\kappa_\varepsilon[B(e_i^\varepsilon) - D(\textstyle\sum_{j\in N} e_i^\varepsilon)]$$
$$+(1-\kappa_\varepsilon)[B(e_i^\varepsilon) - D(\textstyle\sum_{j\in C} e_i^\varepsilon + \sum_{k\in F} e_k^\varepsilon)]\} \quad i \in N \qquad (14)$$

In (14) the two terms in the first curly brackets are the payoffs country *i* receives if it acts as a perfect Kantian with respect to membership, so assumes all other countries have made the same membership decision as country *i*. The first of these two terms, with weight κ_ε, is the payoff it receives if it acts as a perfect Kantian with respect to emissions, while the second term, with weight $1-\kappa_\varepsilon$, is the payoff it receives if it chooses emissions in a non-Kantian fashion. The two terms in the second curly bracket are the payoffs it receives if it acts as a non-Kantian with respect to membership. The first of these two terms, with weight κ_ε, is the payoff it receives if it acts as a perfect Kantian with respect to emissions. The second term, with weight $1-\kappa_\varepsilon$ is just the payoff country *i* would receive in the standard non-cooperative game with self-interested countries, depending whether it is in set *C* or set *F*.

We now solve the game.

2.3.2.1 Stage 2- Equilibrium Emissions

We solve first for the equilibrium emissions e_i^{μ}, when countries act as perfect Kantians with respect to membership, so $\kappa_{\mu} = 1$. From (14), a coalition country *i* takes as given $e_j^{\mu},\ j \neq i$, and chooses e_i^{μ} to maximise:

$$\kappa_{\varepsilon}\{B(e_i^{\mu}) - D(ne_i^{\mu}) + \textstyle\sum_{j\in N_{-i}}[B(e_j^{\mu}) - D(ne_j^{\mu})]\} + (1-\kappa_{\varepsilon})\{\textstyle\sum_{j\in N}[B(e_j^{\mu}) - D(\sum_{k\in N} e_k^{\mu})]\} \quad i \in C \qquad (15a)$$

for which the first-order condition is:

$$B'(e_i^{\mu}) = \kappa_{\varepsilon} n D'(ne_i^{\mu}) + (1-\kappa_{\varepsilon}) \textstyle\sum_{j\in N} D'(\sum_{k\in N} e_k^{\mu}) \quad i \in C \qquad (15b)$$

Imposing symmetry, equilibrium coalition emissions, $\tilde{e}_{\mu}^{c} \equiv \tilde{e}_{\mu}^{c}(n; \kappa_{\varepsilon}, 1)$ solve:

$$B'(\tilde{e}_{\mu}^{c}) = nD'(n\tilde{e}_{\mu}^{c}) \Longrightarrow \tilde{e}_{\mu}^{c} = e^{SO} \qquad i \in C \qquad (15c)$$

so, as we noted at the start of Section 2, when countries act as if all are in coalition, they set the fully cooperative, socially optimal level of emissions.

A fringe country takes as given $e_j^{\mu},\ j \neq i$, and chooses e_i^{μ} to maximise:

$$\kappa_{\varepsilon}[B(e_i^{\mu}) - D(ne_i^{\mu})] + (1-\kappa_{\varepsilon})[B(e_i^{\mu}) - D(\textstyle\sum_{j\in N} e_j^{\mu})] \quad i \in F \qquad (15d)$$

for which the first-order condition is:

$$B'(e_i^{\mu}) = \kappa_{\varepsilon} n D'(ne_i^{\mu}) + (1-\kappa_{\varepsilon}) D'(\textstyle\sum_{k\in N} e_k^{\mu}) \quad i \in F \qquad (15e)$$

Imposing symmetry, equilibrium fringe emissions, $\tilde{e}_{\mu}^{f} \equiv \tilde{e}_{\mu}^{f}(1; \kappa_{\varepsilon}, 1)$ solve:

$$B'(\tilde{e}_{\mu}^{f}) = [\kappa_{\varepsilon} n + (1-\kappa_{\varepsilon})] D'(n\tilde{e}_{\mu}^{f}) \Longrightarrow \tilde{e}_{\mu}^{f} = e^{*}(\kappa_{\varepsilon}) \qquad (15f)$$

so, as we noted in Section 2.1, when countries act as if all countries are in the fringe, they set emissions equal to those in the non-cooperative Kantian equilibrium with Kantian weight κ_{ε}.

We now solve for the equilibrium emissions e_i^{ε}, when countries act as non-Kantians with respect to membership, so $\kappa_{\mu} = 0$. We deal first with the case where both *C* and *F* are non-empty (i.e. $2 \leq m \leq n-1$). From (14), a coalition country *i* takes as given $e_j^{\mu},\ j \neq i$, and chooses e_i^{ε} to maximise:

$$\kappa_{\varepsilon}\{B(e_i^{\varepsilon}) - D(ne_i^{\varepsilon}) + \textstyle\sum_{j\in N_{-i}}[B(e_j^{\varepsilon}) - D(ne_j^{\varepsilon})]\} + (1-\kappa_{\varepsilon})\{\textstyle\sum_{j\in N}[B(e_j^{\varepsilon}) - D(\sum_{k\in C} e_k^{\varepsilon} + \sum_{l\in F} e_l^{\varepsilon})]\} \quad i \in C \qquad (16a)$$

for which the first-order condition is:

$$B'(e_i^{\varepsilon}) = \kappa_{\varepsilon} n D'(ne_i^{\varepsilon}) + (1-\kappa_{\varepsilon}) \textstyle\sum_{j\in C} D'(\sum_{k\in C} e_k^{e} + \sum_{l\in F} e_l^{\varepsilon}) \quad i \in C \qquad (16b)$$

A fringe country takes as given $e_j^{\varepsilon},\ j \neq i$, and chooses e_i^{ε} to maximise:

$$\kappa_\varepsilon[B(e_i^\varepsilon) - D(ne_i^\varepsilon)] + (1-\kappa_\varepsilon)[B(e_i^\varepsilon) - D(\sum_{k\in C} e_k^\varepsilon + \sum_{l\in F} e_l^\varepsilon)] \quad i \in F \qquad (16c)$$

for which the first-order condition is:

$$B'(e_i^\mu) = \kappa_\varepsilon n D'(ne_i^\mu) + (1-\kappa_\varepsilon)D'(\sum_{k\in C} e_k^\mu + \sum_{l\in F} e_l^\varepsilon) \quad i \in F \qquad (16d)$$

Imposing symmetry, equilibrium coalition and fringe emissions, $\tilde{e}_\varepsilon^c \equiv \tilde{e}_\varepsilon^c(n;\kappa_\varepsilon,0)$, $\tilde{e}_\varepsilon^f \equiv \tilde{e}_\varepsilon^f(1;\kappa_\varepsilon,0)$, solve the simultaneous equations:

$$B'(\tilde{e}_\varepsilon^c) = \kappa_\varepsilon n D'(n\tilde{e}_\varepsilon^c) + (1-\kappa_\varepsilon)mD'\left[m\tilde{e}_\varepsilon^c + (m-n)\tilde{e}_\varepsilon^f\right] \qquad (17a)$$

$$B'(\tilde{e}_\varepsilon^f) = \kappa_\varepsilon n D'(n\tilde{e}_\varepsilon^f) + (1-\kappa_\varepsilon)D'\left[m\tilde{e}_\varepsilon^c + (m-n)\tilde{e}_\varepsilon^f\right] \qquad (17b)$$

It is clear that, when *m* = *n*, equilibrium coalition emissions when countries act as non-Kantians with respect to membership are the same as when they act as perfect Kantians with respect to membership, i.e. $\tilde{e}_\varepsilon^c(n;\kappa_\varepsilon,0) = \tilde{e}_\mu^c(n;\kappa_\varepsilon,1) = e^{SO}$, the socially optimal level of emissions. Similarly, when *m* = 1, equilibrium fringe emissions when countries act as non-Kantians with respect to membership are the same as when they act as perfect Kantians with respect to membership, i.e. $\tilde{e}_\varepsilon^f(1;\kappa_\varepsilon,0) = \tilde{e}_\mu^f(1;\kappa_\varepsilon,1)$. So, as we noted above, *in equilibrium*, the distinction between emissions related to Kantian behaviour with respect to emissions and membership is effectively redundant, so that for all values of *m* = 1,…,*n*, we will refer simply to equilibrium Stage 2 emissions: $\tilde{e}^c(m,\kappa_\varepsilon), \tilde{e}^f(m,\kappa_\varepsilon)$ for coalition and fringe countries. The associated equilibrium payoff functions are: $\tilde{\Pi}^c(n,\kappa_\varepsilon) = B[\tilde{e}^c(n,\kappa_\varepsilon)] - D[n\tilde{e}^c(n,\kappa_\varepsilon)] = W^{SO}$; for $2 \le m \le n-1$, $\tilde{\Pi}_\varepsilon^j(m,\kappa_\varepsilon) = B\left[\tilde{e}^j(.)\right] - \kappa_\varepsilon D\left[n\tilde{e}^j(.)\right] - (1-\kappa_\varepsilon)D[m\tilde{e}^c(.) + (n-m)\tilde{e}^f(.)]$, $j = c, f$; $\tilde{\Pi}^f(1,\kappa_\varepsilon) = B[\tilde{e}^f(1,\kappa_\varepsilon)] - D[n\tilde{e}^f(1,\kappa_\varepsilon)] = W^*(\kappa_\varepsilon)$.

The overall equilibrium payoff functions from Stage 2 are given by:

$$\tilde{\Pi}^c(m,\kappa_\varepsilon,\kappa_\mu) = \kappa_\mu\tilde{\Pi}^c(n,\kappa_\varepsilon) + (1-\kappa_\mu)\tilde{\Pi}^c(m,\kappa_\varepsilon) \qquad (18a)$$

$$\tilde{\Pi}^f(m,\kappa_\varepsilon,\kappa_\mu) = \kappa_\mu\tilde{\Pi}^f(1,\kappa_\varepsilon) + (1-\kappa_\mu)\tilde{\Pi}^f(m,\kappa_\varepsilon) \qquad (18b)$$

We know that:

$$\tilde{\Pi}^c(m,1) = \tilde{\Pi}^f(m,1) = W^{SO},\ 1 \le m \le n; \qquad (19a)$$

$$\tilde{\Pi}^c(n,\kappa_\varepsilon) = W^{SO} > \tilde{\Pi}^f(1,\kappa_\varepsilon) = W^*(\kappa_\varepsilon), 0 \le \kappa_\varepsilon < 1; \qquad (19b)$$

$$\tilde{\Pi}^c(m,\kappa_\varepsilon) < \tilde{\Pi}^f(m,\kappa_\varepsilon), 0 \le \kappa_\varepsilon < 1; \qquad (19c)$$

(19a) says that when countries are perfect Kantians with respect to emissions, $\kappa_\varepsilon = 1$, then, no matter what size the coalition is, fringe and coalition countries get the social optimum payoff, because they generate the social optimum emissions. When $\kappa_\varepsilon < 1$, (19b) says that a country is better off when all, or act as if all, are in the grand coalition compared to when all are, or act as if all, are in the fringe, though this gap narrows as κ_ε increases. (19c) is just the standard free-rider benefit that fringe countries derive in the conventional model of IEAs, when not acting as Kantians with respect to membership. The key point of our model of Kantian IEAs is that, with weight κ_μ, this free-riding effect is offset by the benefit coalition countries derive

from asking what benefit would they get if, *hypothetically*, all countries were in the coalition *and set the appropriate emissions*, compared to the benefit fringe countries would get if, *hypothetically*, all countries were in the fringe *and set the appropriate emissions*.

2.3.2.2 Stage 1: Equilibrium Membership

We now determine, for all values of $\kappa_\varepsilon, \kappa_\mu$, the size of the equilibrium membership, $\tilde{m} = \tilde{m}(\kappa_\varepsilon, \kappa_\mu)$. The stability function is:

$$\sigma(m, \kappa_\varepsilon, \kappa_\mu) = \tilde{\Pi}^c(m, \kappa_\varepsilon, \kappa_\mu) - \tilde{\Pi}^f(m-1, \kappa_\varepsilon, \kappa_\mu) \qquad (20)$$

A coalition of size $\tilde{m}, 2 \leq \tilde{m} \leq n-1$ is stable if $\sigma(\tilde{m}, \kappa_\varepsilon, \kappa_\mu) \geq 0$, and $\sigma(\tilde{m}+1, \kappa_\varepsilon, \kappa_\mu) \leq 0$; the grand coalition is stable if $\sigma(n, \kappa_\varepsilon, \kappa_\mu) \geq 0$.

With the general functional forms we have used in Section 2, it is not possible to derive general results about the existence of an equilibrium coalition, whether it is unique, and how big it might be. However, we can illustrate one important implication of our model by writing the stability function as:

$$\begin{aligned}\sigma(m, \kappa_\varepsilon, \kappa_\mu) &= \kappa_\mu\left[\tilde{\Pi}^c(n, \kappa_\varepsilon) - \tilde{\Pi}^f(1, \kappa_\varepsilon)\right] + (1-\kappa_\mu)[\tilde{\Pi}^c(m, \kappa_\varepsilon) - \tilde{\Pi}^f(m-1, \kappa_\varepsilon)] \\ &= \kappa_\mu[W^{SO} - W^*(\kappa_\varepsilon)] + (1-\kappa_\mu)[\tilde{\Pi}^c(m, \kappa_\varepsilon) - \tilde{\Pi}^f(m-1, \kappa_\varepsilon)] \qquad (21)\end{aligned}$$

The second term on the RHS of (21) is the standard stability function when countries act as non-Kantians, and, as we argued in 2.2.1, for a class of commonly employed functional forms for benefits and damage cost functions, becomes negative for values of *m* greater than a small number, typically no greater than 4. The first term in square brackets is the difference in welfare between the fully-co-operative equilibrium and the non-cooperative Kantian equilibrium with Kantian weight κ_ε. It captures the difference between the *full* benefit of joining the coalition and the *full* benefit of joining the fringe, where, by *full* benefit we mean the benefit of assuming, hypothetically, that all countries make the same membership decision *and* set their emissions consistent with that assumption. As κ_μ increases more weight is given to the positive first term and less to the possibly negative second term, suggesting that the size of the stable IEA should increase. The one specific result we can derive, from (19a), is that, when $\kappa_\varepsilon = 1$, both terms in (21) are zero, so any coalition is stable, and we select the grand coalition as the equilibrium.

Result 3

For our general model of IEAs with Kantian behaviour towards emissions and membership, when countries are perfect Kantians with respect to emissions ($\kappa_\varepsilon = 1$), then, for any possible value of the Kantian weight on membership (κ_μ), the stable IEA is the grand coalition, achieving the social optimum level of welfare, W^{SO}.

2.3.2.3 Overall Equilibrium.

Having derived equilibrium membership, $\tilde{m}(\kappa_\varepsilon, \kappa_\mu)$ of the IEA, we can derive the overall equilibrium emissions of a coalition or fringe country (*i* = *c*, *f*):

$$\tilde{e}^i(\kappa_\varepsilon, \kappa_\mu) = \tilde{e}^i[\tilde{m}(\kappa_\varepsilon, \kappa_\mu), \kappa_\varepsilon], \qquad (22a)$$

and the overall welfare of a coalition or fringe country:

$$\widetilde{W}^i(\kappa_\varepsilon, \kappa_\mu) = B\left[\tilde{e}^i(\kappa_\varepsilon, \kappa_\mu)\right] - D\{\tilde{m}(\kappa_\varepsilon, \kappa_\mu)\tilde{e}^c(\kappa_\varepsilon, \kappa_\mu) + \left[n - \tilde{m}(\kappa_\varepsilon, \kappa_\mu)\right]\tilde{e}^f(\kappa_\varepsilon, \kappa_\mu)\} \quad (22b)$$

As we have noted, to derive further results for our model, we need to turn to specific functional forms for the benefit and damage cost functions.

3 Results for Specific Functional Forms.

Eichner and Pethig (2022) derive more specific numerical results concerning equilibrium IEA membership by employing the special case where the benefit function is quadratic and the damage cost function is linear. In the Online Appendix B to this paper, we present results for our model using the same functional forms. In this paper, we employ a broader set of functional forms by assuming that the damage cost function is also quadratic. In Section 3.1 we present a range of results for our model concerning Stage 2 equilibrium emissions, Stage 1 equilibrium IEA membership, and a range of other significant outcomes. In Section 3.2 we compare the key results from our model with those derived from the model of Eichner and Pethig (2022), and show that, on 3 key metrics, the results from our model outperform those from the model of Eichner and Pethig.

In our numerical results we will assume that the number of countries, *n*, equals 100 (which we take to be a 'large' number of countries), but employ a wide range of values of the key parameters $(\kappa_\varepsilon, \kappa_\mu)$. The results address the following four questions.

(1) How do Stage 2 equilibrium emissions $\tilde{e}^c(m, \kappa_\varepsilon), \tilde{e}^f(m, \kappa_\varepsilon)$ vary as *m*, κ_ε vary?
(2) What is the size of the equilibrium coalition? In particular, for what values of $(\kappa_\varepsilon, \kappa_\mu)$ is the equilibrium the grand coalition?
(3) For given values of $(\kappa_\varepsilon, \kappa_\mu)$, to what extent does our model of IEAs with imperfect Kantian behaviour close the gaps in emissions and welfare between the non-cooperative and social optimum outcomes for a coalition country, a fringe country, for all countries?
(4) Aggregating across all countries, what is the relative contribution of Kantian behaviour and IEA formation to closing the gaps in emissions and welfare between the non-cooperative non-Kantian equilibrium and first-best. This addresses the question raised by Nowakowski and Oswald (2020) of what weight policy-makers should give to trying to influence individuals to make consumption choices which have lower carbon footprints compared to trying to persuade national governments to join an IEA.

To address questions (3) and (4), it will be useful to construct measures of equilibrium emissions and welfare of an *average* country defined as:

$$\tilde{e}^a(\kappa_\varepsilon, \kappa_\mu) \equiv \{\tilde{m}(\kappa_e, \kappa_\mu)\tilde{e}^c(\kappa_\varepsilon, \kappa_\mu) + [n - \tilde{m}(\kappa_\varepsilon, \kappa_\mu)]\tilde{e}^f(\kappa_\varepsilon, \kappa_\mu)\}/n \qquad (23a)$$

$$\widetilde{W}^a(\kappa_\varepsilon, \kappa_\mu) \equiv \{\tilde{m}(\kappa_e, \kappa_\mu)\widetilde{W}^c(\kappa_\varepsilon, \kappa_\mu) + [n - \tilde{m}(\kappa_\varepsilon, \kappa_\mu)]\widetilde{W}^f(\kappa_\varepsilon, \kappa_\mu)\}/n \qquad (23b)$$

It will also be useful to construct measures of how far non-cooperative Kantian behaviour, or the formation of an IEA with different degrees of Kantian weights, closes the gaps between

non-cooperative non-Kantian emissions and welfare $[e^{NC} = \hat{e}(0), W^{NC} = \widehat{W}(0)]$ and social optimal emissions and welfare $[e^{SO} = \hat{e}(1), W^{SO} = \widehat{W}(1)]$. We define these measures as:

$$\hat{\hat{e}}(\kappa) \equiv \frac{e^{NC} - \hat{e}(\kappa)}{e^{NC} - e^{SO}}; \quad \widehat{\widehat{W}}(\kappa) \equiv \frac{\widehat{W}(\kappa) - W^{NC}}{W^{SO} - W^{NC}} \tag{23c}$$

for the non-cooperative Kantian equilibrium, and

$$\tilde{\tilde{e}}^{j}(\kappa_\varepsilon, \kappa_\mu) \equiv \frac{e^{NC} - \tilde{e}^{j}(\kappa_\varepsilon, \kappa_\mu)}{e^{NC} - e^{SO}}; \quad \widetilde{\widetilde{W}}^{j}(\kappa_\varepsilon, \kappa_\mu) \equiv \frac{\widetilde{W}^{j}(\kappa_\varepsilon, \kappa_\mu) - W^{NC}}{W^{SO} - W^{NC}} \quad j = c, f, a \tag{23d}$$

for the equilibrium of an IEA.

As we will see, another benefit of employing these measures of emissions and welfare gaps in is that they are less dependent on some of parameters used in our specific functional forms.

3.1 Results with Quadratic Benefit Function and Quadratic Damage Cost Function

We assume that the benefit function takes the form $B(e_i) = \beta e_i - 0.5e_i^2$ and the damage cost function takes the form: $D[\sum_{j\in N} e_j] = 0.5\delta[\sum_{j\in N} e_j]^2$. To ensure non-negative emissions for coalition countries, we show in Appendix A that the damage cost parameter must satisfy: $\delta < \overline{\delta} \equiv 4/(n-1)^2$. Appendix A presents the key results for the first-best, non-cooperative self-interested equilibrium, the non-cooperative equilibrium with Kantian behaviour, and our model of IEAs with Kantian behaviour towards emissions and membership. We are able to obtain analytical results only for equilibrium Stage 2 emissions; so, the remaining results in this sub-section are derived numerically. Many of the key results we derived from the model with linear damage costs still apply with quadratic damage costs, so we will focus on the additional results that arise with quadratic damage costs.

3.1.1. Question (1): Stage 2 Equilibrium Emissions.

For $m = 1$, all countries are in the fringe, so $\tilde{e}^{f}(1, \kappa_\varepsilon) = \hat{e}(\kappa_\varepsilon)$; for $m = n$, all countries are in the (grand) coalition, so $\tilde{e}^{c}(n, \kappa_\varepsilon) = e^{SO}$. For $2 \le m \le n-1$, we first define: $\phi(\kappa_\varepsilon) = \frac{1+\delta\kappa_\varepsilon n^2}{(1-\kappa_\varepsilon)\delta}$.

Equilibrium Stage 2 emissions are:

$$\tilde{e}^{c}(m, \kappa_\varepsilon) = \frac{\beta[\phi(\kappa_\varepsilon) - (n-m)(m-1)]}{(1+\delta\kappa_\varepsilon n^2)[\phi(\kappa_\varepsilon) + (n+m^2-m)]} > 0; \tag{24a}$$

$$\tilde{e}^{f}(m, \kappa_\varepsilon) = \frac{\beta[\phi(\kappa_\varepsilon) + m^2 - m]}{(1+\delta\kappa_\varepsilon n^2)[\phi(\kappa_\varepsilon) + (n+m^2-m)]} > 0; \tag{24b}$$

$$\begin{aligned} \tilde{e}^{a}(m, \kappa_\varepsilon) &= [m * \tilde{e}^{c}(m, \kappa_\varepsilon) + (n-m)\tilde{e}^{f}(m, \kappa_\varepsilon)]/n \\ &= \frac{\beta n \phi(\kappa_\varepsilon)}{(1+\delta\kappa_\varepsilon n^2)[\phi(\kappa_\varepsilon) + (n+m^2-m)]} \end{aligned} \tag{24c}$$

It is straightforward to see that $\tilde{e}^{f}(m, \kappa_\varepsilon) > \tilde{e}^{c}(m, \kappa_\varepsilon), \ 2 \le m \le n-1$.

We first consider how equilibrium emissions vary as the size of the IEA, m, varies. The interesting case is the behaviour of $\tilde{e}^{c}(m, \kappa_\varepsilon)$. In Appendix A we define a key value of $\kappa_\varepsilon, \bar{\kappa}_\varepsilon, 0 < \bar{\kappa}_\varepsilon < 1$, and, for, $0 \le \kappa_\varepsilon < \bar{\kappa}_\varepsilon$, two key values of m, $\underline{m}(\kappa_\varepsilon)$, $\hat{m}(\kappa_\varepsilon)$, where $0 < \underline{m}(\kappa_\varepsilon) < \hat{m}(\kappa_\varepsilon) < n$. Then we have:

Result 4(a)

Varying membership affects Stage 2 equilibrium emissions as follows[21]*:*

(i) $\frac{\partial \tilde{e}^a}{\partial m} < 0 \quad \forall m, \kappa$

(ii) $\frac{\partial \tilde{e}^f}{\partial m} > 0 \quad \forall m, \kappa$

(iii) $For\ 1 > \kappa_\varepsilon \geq \bar{\kappa}_\varepsilon, \frac{\partial \tilde{e}^c}{\partial m} < 0,\ \forall m,\ 2 \leq m \leq n-1$

(iv) $For\ 0 < \kappa_\varepsilon < \bar{\kappa}_\varepsilon\ \frac{\partial \tilde{e}^c}{\partial m} \leq 0 \iff 2 \leq m < \hat{m}(\kappa_\varepsilon);\ \frac{\partial \tilde{e}^c}{\partial m} > 0 \iff \hat{m}(\kappa_\varepsilon) \leq m \leq n.$

(v) $For\ 0 < \kappa_\varepsilon < \bar{\kappa}_\varepsilon\ \tilde{e}^c(m, \kappa_\varepsilon) \leq e^{SO} \iff \underline{m}(\kappa_\varepsilon) \leq m \leq n$

As we would expect, average (and hence total) emissions fall as the IEA gets larger. However, each fringe country *increases* its emissions[22], essentially because the incentive to free ride gets bigger as the coalition gets. As noted, the interesting case is the effect of rising membership on $\tilde{e}^c(m, \kappa)$. If $\kappa_\varepsilon \geq \bar{\kappa}_\varepsilon$ then, as membership increases, each coalition country cuts its emissions from the non-cooperative level to the socially optimal level, both because it places a high Kantian weight on emissions, and because that ensures that average emissions fall despite each fringe country increasing its emissions. However, if $\kappa_\varepsilon < \bar{\kappa}_\varepsilon$, then the cut in emissions by each coalition country as membership rises now requires that when membership reaches a critical level, $\underline{m}(\kappa_\varepsilon)$, coalition countries cut their emissions *below* the socially optimal level, and this continues until each coalition country's emissions reach a minimum when membership is $\hat{m}(\kappa_\varepsilon)$, and then each coalition country increases its emissions to ensure that its emissions reach the socially optimal level when the grand coalition forms. This effect is larger the *lower* is the value of κ_ε.

To get a feel for the values of $\bar{\kappa}_\varepsilon$, $\hat{m}(\kappa_\varepsilon)$ that arise in the analysis of $\frac{\partial \tilde{e}^c}{\partial m}$, we take parameters *n* = 100 (i.e. a 'large' number of countries), and $\delta = 0.9\bar{\delta}$, (i.e. damage costs are large). From equation (A5) in Appendix A, we calculate $\bar{\kappa}_\varepsilon = 0.3542$. In Table 1(a) for a range of values of $\kappa_\varepsilon < \bar{\kappa}_\varepsilon$ we present the corresponding values of $\underline{m}(\kappa_\varepsilon), \hat{m}(\kappa_\varepsilon)$, and $\hat{\tilde{e}}^c[\hat{m}(\kappa_\varepsilon), \kappa_\varepsilon]$, which measures the maximum extent to which a coalition country cuts its emissions from the non-cooperative equilibrium relative to the gap needed to reduce emissions to the socially optimal level; a value greater than 1 indicates that it cuts emissions to a level below the social optimum.

Thus, for $\kappa = 0.15$, for example, when membership increases from 2 to 48, a coalition country's emissions fall, but are still above the social optimum of emissions. When membership rises above 48 a coalition country's emissions continue to fall, but are now below the socially optimal level. Coalition country's emissions continue to fall until membership reaches 69, at which point the cut in emissions has reached 3.89% above the level needed to reach the socially optimal level of emissions. When membership rises above 69, a coalition county's emissions start to rise until they reach the socially optimal level when the grand coalition forms. When $\kappa > \bar{\kappa} = 0.3615$, coalition emissions fall steadily as membership rises, reaching the social level when the grand coalition forms.

[21] We recognise that the partial derivatives treat *m* as a real variable when it is an integer variable. This can be problematic if *n* is small, but we assume that it is large. Moreover, the key results are supported by numerical simulations.

[22] As we show in the Online Appendix B, in the special case with linear damage costs, fringe country emissions are independent of the size of the coalition.

κ_ε	$\underline{m}(\kappa_\varepsilon)$	$\widehat{m}(\kappa_\varepsilon)$	$\hat{\tilde{e}}^c[\widehat{m}(\kappa_\varepsilon), \kappa_\varepsilon]$
0.010	29	52	1.2116
0.050	34	57	1.1317
0.100	41	63	1.0726
0.150	49	69	1.0389
0.200	59	76	1.0197
0.250	69	82	1.0090
0.300	81	89	1.0036
0.350	95	97	1.0015

Table 1(a): Effects on Equilibrium Stage 2 Coalition Emissions of Variations in Membership

This raises the question of how it can be the case that average emissions can be falling throughout as membership rises, when, for some parameter values, *both* coalition and fringe countries are increasing their emissions. The answer is that coalition emissions are well below fringe emissions, so as membership increases, this can still be compatible with average emissions falling.

We now turn to the effects of variation in the Kantian weight κ_ε on equilibrium emissions. In Appendix A, for values of *m*, $2 \leq m \leq n$ we define a key value of $\kappa_\varepsilon, \check{\kappa}_\varepsilon(m)$, and prove the following result.

Result 4(b)

The effect of variations in the Kantian weight on emissions, κ_ε, on Stage 2 equilibrium emissions are as follows:

(i) $\frac{\partial \tilde{e}^a}{\partial \kappa_\varepsilon} \leq 0 \iff m \leq n, \ \forall \kappa_\varepsilon$

(ii) $\frac{\partial \tilde{e}^f}{\partial \kappa_\varepsilon} < 0 \quad \forall m, \kappa_\varepsilon$

(iii) If $\check{\kappa}_e(m) \leq 0, \frac{\partial \tilde{e}^c}{\partial \kappa_\varepsilon} < 0 \ \forall \kappa_\varepsilon$; if $\check{\kappa}_\varepsilon(m) > 0, \frac{\partial \tilde{e}^c}{\partial \kappa_\varepsilon} \leq 0 \iff 1 \geq \kappa_\varepsilon \geq \check{\kappa}_\varepsilon(m)$

As with Result 4(a), average, and hence total, emissions fall as the Kantian weight increases. This also applies to emissions of a fringe country, and, if $\check{\kappa}_e(m)$ is not positive, the emissions of a coalition country[23]. However, when $\check{\kappa}_e(m)$ is positive, emissions of a coalition country increase as κ_ε increases for values of $\kappa_\varepsilon < \check{\kappa}_e(m)$, and then fall when $\kappa_\varepsilon \geq \check{\kappa}_e(m)$.The intuition behind this result stems from the result that coalition emissions are always less than fringe emissions. When *m* is small, there is not much scope for free riding by the fringe, so the gap between coalition and fringe emissions is small. Since overall emissions and fringe emissions fall as κ increases, coalition emissions must also fall. However, when *m* becomes larger, there is more scope for free riding by fringe countries, so the gap between emissions of a fringe coalition country gets larger, giving scope for coalition countries to increase their emissions, but this can only go so far because fringe emissions are falling as κ increases, reducing the scope for coalition emissions to increase while remaining below fringe emissions, so coalition emissions must fall as κ increases beyond this critical value.

[23] In the Online Appendix B we show that these results also apply to the special case with linear damage costs.

To get a feel for the values of $\check{\kappa}(m)$ we again take the case where n =100, $\delta = 0.95\bar{\delta}$. Table 2 presents values of $\check{\kappa}(m)$ for a range of values of m between 2 and 99.

,

m	$\check{\kappa}(m)$	m	$\check{\kappa}(m)$	m	$\check{\kappa}(m)$	m	$\check{\kappa}(m)$
2	-1.9969	20	0.2594	50	0.6883	80	0.8095
5	-1.0056	25	0.3925	55	0.7172	85	0.8216
10	-0.2879	30	0.4866	60	0.7415	90	0.8324
13	-0.0570	35	0.5565	65	0.7623	95	0.8420
14	0.0031	40	0.6105	70	0.7802	97	0.8455
15	0.0565	45	0.6534	75	0.7958	99	0.8490

Table 1(b): Values of $\check{\kappa}(m)$ for Different Values of m

Thus, when m lies below 14, coalition emissions always fall as κ increases. For larger values of m, there is a critical positive value of κ such that coalition emissions increase if κ lies below that critical value, and this critical rises from about 0.05 to 0.85 as m increases from 15 to 99.

To summarise: (i) average (total) emissions fall as *both* membership and Kantian weight on emissions increase, so, seeking to form a larger coalition and increasing Kantian weight on moral behaviour are *substitute* approaches to cutting emissions at Stage 2; (ii) a fringe country's emissions rise as membership increases but fall as the Kantian weight on emissions rises; (iii) a coalition country's emissions rise and fall as both membership and the Kantian weight on emissions increase, depending on precise parameter values. Interestingly, there are parameter values for which the emissions of both a fringe country and a coalition country increase as membership increases, although average emissions fall.

3.1.2 Question (2): Size of Equilibrium IEA

In the Online Appendix, we show that with a linear damage cost function, the stability function is effectively independent of δ and κ_ε, and is decreasing in the size of IEA membership, m, so, for given parameters $\kappa_\varepsilon, \kappa_\mu$ there is a unique stable IEA.

With quadratic damage costs, the stability function does depend on δ and κ_ε, and is initially decreasing and then increasing in m. We illustrate the implications in Figure 2, for n = 100, $\delta = 0.9\bar{\delta}$ and $\kappa_\varepsilon = 0.05$. When $\kappa_\mu = 0.15$, the stability function intersects the horizontal axis once, so there is a unique stable IEA with 26 members. When $\kappa_\mu = 0.23$, the stability function intersects the horizontal axis twice; the lower value results in a stable IEA of size 41; the upper intersection, with m approximately 0.89, does not yield a stable IEA, but the grand coalition is a second stable IEA; we apply an equilibrium selection rule of selection the higher stable IEA, the grand coalition. With $\kappa_\mu = 0.3$, the stability function does not intersect the horizontal axis so the unique stable IEA is the grand coalition.

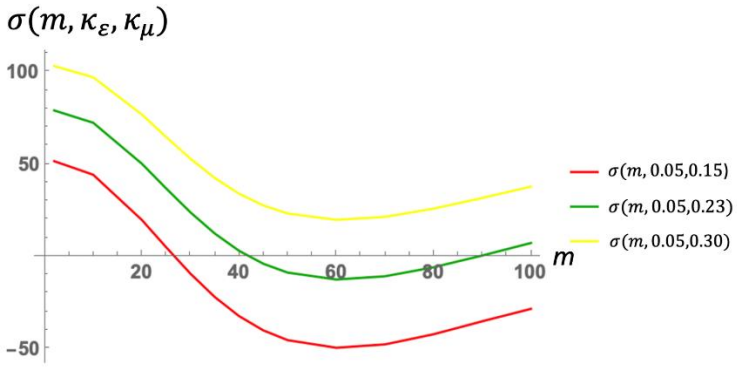

Figure 2: Stability Function for Different Values of κ_μ

δ	$0.9\bar{\delta}$	$0.5\bar{\delta}$
κ_ε	κ_μ	κ_μ
0.00	0.19-0.24	0.26-0.27
0.05	0.22-0.25	0.28
0.10	0.24-0.26	..
0.20	0.28	..

Table 2: Range of values of Kantian weight on membership with two stable IEAs

Table 2 shows the range of values of κ_μ for which we obtain two stable IEAs (the higher being the grand coalition), using two values of $\delta = 0.9\bar{\delta}$, $\delta = 0.5\bar{\delta}$, and values of κ_ε between 0.0 and 0.20. We also used the value of $\delta = 0.25\bar{\delta}$ and the same values of κ_ε, but there were no instances of two stable IEAs. As Table 2 shows, instances of two stable IEAs arise with high values of damage costs and low values of the Kantian weight on emissions.

Result 5

With quadratic benefit function and damage cost function, the stability function is initially decreasing in m and then increasing; for large values of κ_μ and low values of κ_ε there can be two stable IEAs, one relatively small, the other the grand coalition.

We now determine the minimum value of κ_μ needed to secure the grand coalition. In the Online Appendix we show that with linear damage costs this minimum value, $\underline{\kappa_\mu}(n)$, depends only on n, and tends to 0.5 as $n \longrightarrow \infty$, with $\underline{\kappa_\mu}(100) = 0.495$. With quadratic damage costs, this minimum value of κ_μ it also depends on κ_ε and δ, so we define the minimum value as $\underline{\kappa_\mu}(n, \kappa_\varepsilon, \delta)$. In Figure 3 we fix n = 100, and show the values of $\underline{\kappa_\mu}(100, \kappa_\varepsilon, \delta)$ for values of κ_ε between 0.00 and 1.00 and $\delta = 0.9\bar{\delta}$, $0.5\bar{\delta}$, and $0.25\bar{\delta}$. We see that (i) $\underline{\kappa_\mu}(100, \kappa_\varepsilon, \delta)$ is increasing in κ_ε; but, (ii), decreasing in δ. Thus, the grand coalition can now be achieved with much smaller Kantian weights: e.g. $\underline{\kappa_\mu}\left(100,0.0,0.9\overline{\delta}\right) = 0.2$.

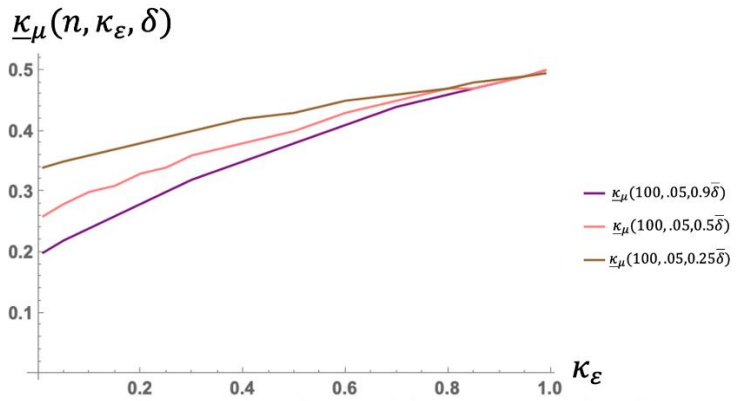

Figure 3. Minimum Value of κ_μ Needed to Secure Grand Coalition

The rationale for (i) is that an increase in κ_ε increases the emissions that all countries would carry out in the absence of an IEA, reducing the conventional benefits of joining an IEA, so a higher weight on membership is needed to offset that effect. The rationale for (ii) is that, with quadratic damage costs, an increase in the level of damage costs significantly increases the benefit of securing the grand coalition, and so requires a lower Kantian weight on membership to achieve that outcome.

Finally, in Figure 4, we again fix n = 100, set $\delta = 0.9\overline{\delta}$, and plot the size of the stable IEA $\widetilde{m}(\kappa_\varepsilon, \kappa_\mu)$ for values of κ_μ between 0 and 0.5 (we know the grand coalition is stable for $\kappa_\mu >$ 0.5), and κ_ε = 0.1, 0.3, and 0.5.

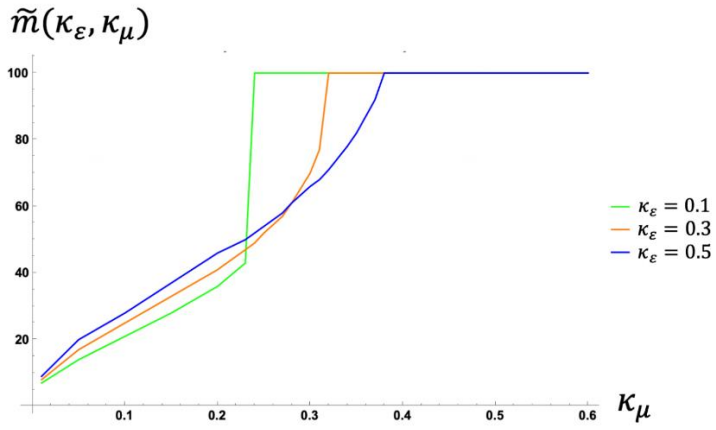

Figure 4: Equilibrium IEA Membership

As expected, for any value of κ_ε, the size of equilibrium IEA membership is increasing in the Kantian weight on membership, κ_μ. But if we compare these curves for two different values of κ_ε, we see that they intersect. There are three factors leading to this result. (i) We know that as $\kappa_\varepsilon, \kappa_\mu \longrightarrow 0, \widetilde{m}(\kappa_\varepsilon, \kappa_\mu) \longrightarrow 2$. (ii) For a given but low value of κ_μ, an increase in κ_ε causes coalition countries to cut their emissions by less fringe countries, since they are already acting to cut emissions, causing the stability function and hence the size of a stable IEA to increase. (iii) Finally, we showed above that $\underline{\kappa}_\mu(n, \kappa_\varepsilon, \delta)$ is decreasing in κ_ε. Putting these three arguments together produces the result in Figure 4, where, for values of $\kappa_\mu < 0.25$, $\widetilde{m}(\kappa_\varepsilon, \kappa_\mu)$ is increasing across the 3 values of κ_ε, while for values of $\kappa_\mu > 0.30$, $\widetilde{m}(\kappa_\varepsilon, \kappa_\mu)$ is decreasing across the 3 values of κ_ε.

Result 6

With quadratic benefit function and damage cost function, the size of the stable IEA, $\widetilde{m}(\kappa_\varepsilon, \kappa_\mu)$ *is increasing in* κ_μ *but for values of* κ_μ *above a critical level, is decreasing in* κ_ε*. The grand coalition is always stable with* $\kappa_\mu \geq 0.5$*, but can be achieved with much smaller values, e.g* $\kappa_\mu = 0.2, \kappa_\varepsilon = 0.0$.

This completes the analysis of the size of the stable IEA and we now turn to the extent to which the stable IEA is able to close the emissions and welfare gaps.

3.1.3 Question (3): Closure of Emissions and Welfare Gaps of Coalition and Fringe Countries

We consider first the closure of emission gaps, starting with the emissions gap for a coalition country. Values of $\tilde{\tilde{e}}^c(\kappa_\varepsilon, \kappa_\mu)$, are plotted in Fig 5a, using the same range of values for $\kappa_\varepsilon, \kappa_\mu$ as for Figure 4.

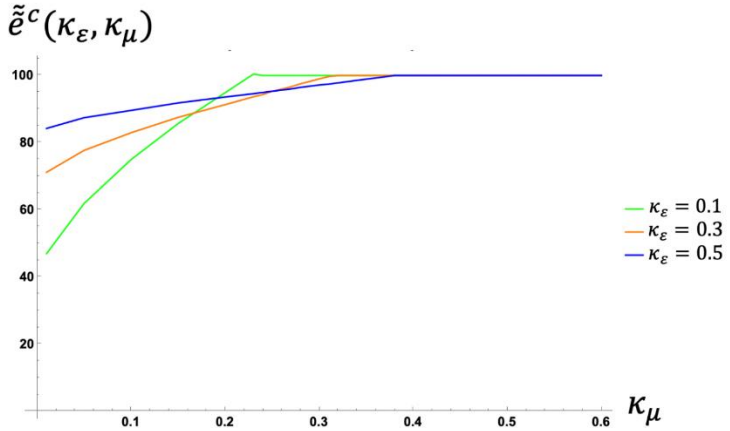

Figure 5a: Equilibrium Emissions of Coalition Country

For low values of κ_μ, the extent to which the coalition emission gap is closed is increasing in the two Kantian weights[24] $\kappa_\varepsilon, \kappa_\mu$. However, there are two important differences for higher values of for higher values of κ_μ. First, it is now possible that $\frac{\partial \tilde{\tilde{e}}^c(.)}{\partial \kappa_\varepsilon} < 0$[25]. The rationale is that, as we saw in Section 3.1.2, the grand coalition is achieved with smaller values of the Kantian weight on membership. Second, as we noted in Section 3.1.1, for sufficiently low values of κ_ε, it is possible that $\tilde{\tilde{e}}^c(.) > 100\%$. This arises in Figure 5a: when κ_ε= 0.1, the grand coalition is achieved when κ_μ=0.24; when κ_μ=0.23, $\tilde{\tilde{e}}^c(0.1, 0.23) = 100.45$. This does not arise for the higher values of κ_ε we have used.

The emissions gap for a fringe country, $\tilde{\tilde{e}}^f(\kappa_\varepsilon, \kappa_\mu)$ is plotted in Figure 5b.

[24] Similar to the case with linear damage costs, as we show in the Online Appendix.
[25] with linear damage costs, $\tilde{\tilde{e}}^c(\kappa_\varepsilon, \kappa_\mu)$ rises steadily towards 100% as the Kantian weight on membership rises.

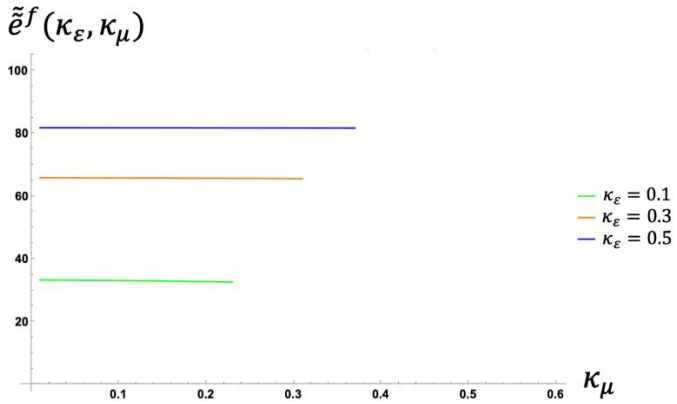

Figure 5b: Equilibrium Emissions of a Fringe Country

It looks as if the extent to which the emissions gap of a fringe country is closed is again independent of κ_μ, though with a value significantly greater than κ_ε. Figure 5c shows the result for κ_ε = 0.3, but with significantly magnified scale on the vertical axis. This shows that $\tilde{\tilde{e}}^f(0.3, \kappa_\mu)$ falls from 65.89 to 65.57 as κ_μ rises from 0.01 to 0.31 (the grand coalition is achieved when κ_μ = 0.32). Thus, the size of the emissions gap of a fringe country increases, slightly, as the Kantian weight on membership increases.

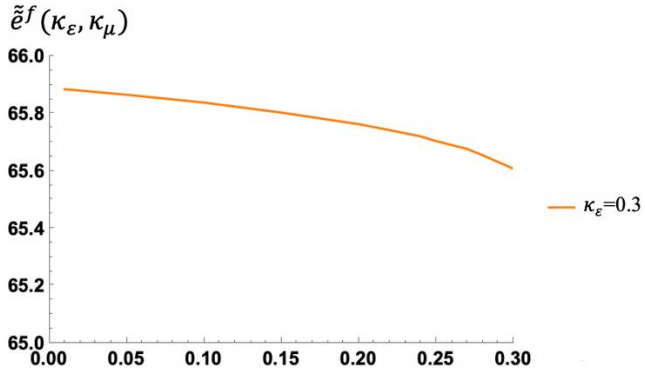

Figure 5c: Equilibrium Emissions of Fringe Country For κ_ε=0.3

We now turn to closure of welfare gaps[26]. Values of $\tilde{\tilde{W}}^c(\kappa_\varepsilon, \kappa_\mu)$ are plotted in Figure 6a. For high values of κ_ε (0.3, 0.5) and low values of κ_μ (< 0.2) coalition welfare can fall as κ_μ increases; in the Online Appendix B, we show that this also occurs when damage costs are linear, and, as we explain there, it arises when the equilibrium membership is below a critical value, which can be the case with small values of κ_μ. Values of $\tilde{\tilde{W}}^f(\kappa_\varepsilon, \kappa_\mu)$ are plotted in Figure 6b. For high values of κ_ε (0.3, 0.5) and κ_μ, $\tilde{\tilde{W}}^f(\kappa_\varepsilon, \kappa_\mu)$ can exceed 100%, though not exceeding 103% for the values in Fig 6a[27]. This

[26] As we show in the Online Appendix B, the main results are similar to the case with linear damage costs.
[27] Again, we show in the Online Appendix B that this can also occur with linear damage costs, though the margin can be much higher, up to 130%.

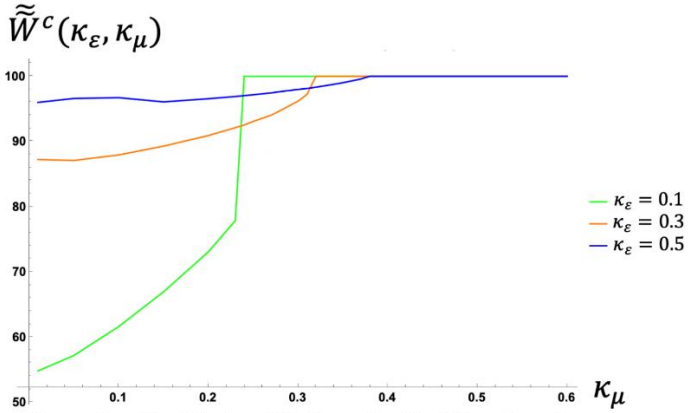

Figure 6a: Equilibrium Welfare of a Coalition Country

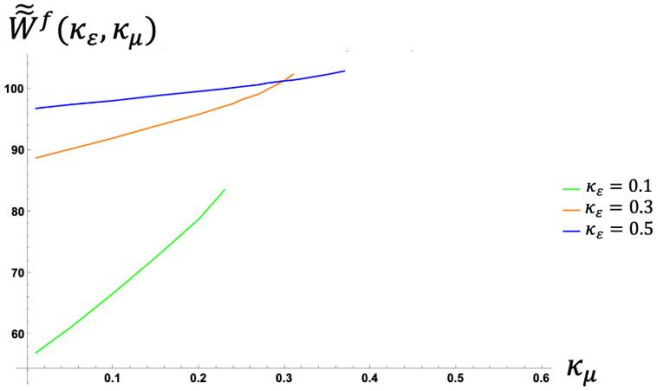

Figure 6b: Equilibrium Welfare of a Fringe Country

We summarise these findings in Result 7.

Result 7

With quadratic benefit function and quadratic damage cost function:

- *(i)* *The emissions gap for a coalition country is decreasing in κ_μ, but, for large values of κ_μ, may be increasing in κ_ε; emissions of a coalition country could fall below first-best for values of κ_μ close to that which secures the grand coalition;*
- *(ii)* *The emissions gap for a fringe country is decreasing in κ_ε but is increasing, slightly, in κ_μ;*
- *(iii)* *The welfare gap for a coalition country decreases with increases in both Kantian weights, except for relatively high values of κ_ε (≥ 0.3) and low values of κ_μ (< 0.2) when it increases in κ_μ;*
- *(iv)* *The welfare gap for a fringe country is decreasing in $\kappa_\varepsilon, \kappa_\mu$; welfare of a fringe country can exceed first-best, though to a smaller extent than with a linear damage cost function.*

3.1.4 Question (4): Extent to which Overall Emissions and Welfare Gaps are Closed by Countries Acting in a Kantian Fashion and by Forming an IEA

Finally, we address the questions of the extent to which the emissions and welfare gaps are closed by countries acting in a Kantian fashion in a non-cooperative equilibrium and what

additional contribution is made by countries seeking to form an IEA. The results are shown in Tables 3a and 3b for emissions and welfare respectively.

The main result is that the non-cooperative equilibrium makes significant contributions to closing the emissions and, particularly, welfare gaps[28]. Thus, the additional contribution of forming an IEA gets squeezed. This is particularly true for welfare where, for all three values of κ_ε, the non-cooperative equilibrium contributes over 50% to closing the welfare gap. Nevertheless, it is still the case that changing moral attitudes and negotiating an IEA are *complementary* approaches to tackling climate change, not substitutes.

κ_ε	Imperfect Kantian Non-Coop Equilibrium	IEAs- κ_μ		
		0.1	0.3	0.5
0.1	33.37	41.95	100.00	100.00
0.3	65.89	70.12	88.97	100.00
0.5	81.84	84.01	91.94	100.00

Table 3a: Contribution of Changing Moral Attitudes and Negotiating IEAs to Reducing Global Emissions

κ_ε	Imperfect Kantian Non-Coop Equilibrium	IEAs- κ_μ		
		0.1	0.3	0.5
0.1	55.61	65.61	100.00	100.00
0.3	88.37	90.99	97.75	100.00
0.5	96.70	97.44	99.13	100.00

Table 3b: Contribution of Changing Moral Attitudes and Negotiating IEAs to Raising Global Welfare

3.2 Comparison of Results From Our Model With Those From Eichner and Pethig (2022)

In this section, we use the model in Section 3.1 with quadratic benefit and damage cost functions to compare results derived from our model of IEAs (denoted UU) with those derived from the model of IEAs in Eichner and Pethig (2022) (denoted EP). To make such comparison we need to choose parameter values for $\kappa_\varepsilon, \kappa_\mu$ which are applicable to both models. EP use three parameters, $\alpha, \kappa_\varepsilon, \kappa_\mu$, where: $0 \leq \alpha \leq 1$; $0 \leq \kappa_\varepsilon \leq 1$; $0 \leq \kappa_\mu \leq 1$. We argued that the parameter α is effectively redundant, so we will work just with the parameters $\kappa_\varepsilon, \kappa_\mu$, but the

[28] We show in the Online Appendix B that the contribution of the non-cooperative Kantian equilibrium to closing the emissions and welfare gaps are smaller with linear damage costs, and hence the contribution of forming an IEA is larger, for all values of the Kantian weights, except $\kappa_\varepsilon = 0.1, \kappa_\mu = 0.3$.

constraint on the value of α is replaced by the constraint that $0 \leq \kappa_\varepsilon + \kappa_\mu \leq 1$. We will choose parameter values that satisfy this constraint.

We choose the high damage cost parameter $\delta = 0.9\bar{\delta}$,$\kappa_\varepsilon = 0.1$, 0.3 and 0.5, with κ_μ ranging between 0.05 and 0.4[29]. We focus on 3 key measures of performance of these models: the size of the equilibrium IEA, the % of the emissions gap closed by the equilibrium IEA and the % of the welfare gap closed by the equilibrium IEA. The results are shown in Tables 4a, 4b, 4c.

κ_μ	Size of Equilibrium IEA					
	$\kappa_\varepsilon = 0.1$		$\kappa_\varepsilon = 0.3$		$\kappa_\varepsilon = 0.5$	
	UU	EP	UU	EP	UU	EP
0.05	14	10	17	12	20	11
0.10	21	15	25	16	28	13
0.15	28	22	33	20	37	14
0.20	36	32	41	27	45	16
0.25	100	100	53	38	54	21
0.30	100	100	70	63	66	30
0.35	100	100	100	100	82	54
0.40	100	100	100	100	100	100

Table 4a: Comparison of Size of Equilibrium IEA in Models UU and EP

κ_μ	% of Emissions Gap Closed by IEA					
	$\kappa_\varepsilon = 0.1$		$\kappa_\varepsilon = 0.3$		$\kappa_\varepsilon = 0.5$	
	UU	EP	UU	EP	UU	EP
0.05	37.29	36.10	67.88	67.32	82.96	82.43
0.10	41.95	39.41	70.12	68.62	84.01	82.85
0.15	47.78	44.67	72.98	70.12	85.53	83.20
0.20	55.18	52.77	76.25	72.71	87.12	83.69
0.25	100.00	100.00	81.09	76.94	89.10	84.66
0.30	100.00	100.00	88.97	86.70	91.94	86.39
0.35	100.00	100.00	100.00	100.00	95.81	91.04
0.40	100.00	100.00	100.00	100.00	100.00	100.00

Table 4b: Comparison of % of Emissions Gap Closed in Models UU and EP

[29] We did the calculations also for the values of κ_μ = 0.45 and 0.5, but the results are the same as for $\kappa_\mu = 0.4$ so we do not present them here.

κ_μ	% of Welfare Gap Closed by IEA					
	$\kappa_\varepsilon = 0.1$		$\kappa_\varepsilon = 0.3$		$\kappa_\varepsilon = 0.5$	
	UU	EP	UU	EP	UU	EP
0.05	60.53	59.07	89.68	89.31	97.10	96.91
0.10	65.61	62.77	90.99	90.08	97.44	97.05
0.15	71.01	67.93	92.43	90.88	97.87	97.16
0.20	76.73	74.70	93.84	92.12	98.25	97.29
0.25	100.00	100.00	95.57	93.89	98.66	97.54
0.30	100.00	100.00	97.75	97.08	99.13	97.94
0.35	100.00	100.00	100.00	100.00	99.62	98.85
0.40	100.00	100.00	100.00	100.00	100.00	100.00

Table 4c: Comparison of % of Welfare Gap Closed in Models UU and EP

We see that, for all three measures of the effectiveness of IEAs, and for all values of the Kantian weights, the model employed in this paper (UU) are greater than for the model employed in Eichner and Pethig (2022). We believe that this is because our model captures the full benefit of deciding to be a member of an IEA compared to being a member of the fringe by choosing emissions appropriate to those decisions.

4 Conclusions

This paper contributes to a literature which seeks to overcome the pessimistic conclusion drawn from standard non-cooperative two-stage game models of IEAs that the number of countries who join an IEA is small precisely when the potential gains from achieving the grand coalition are large. In the standard model, agents act in a self-interested manner. We draw on the important work of Alger and Weibull (2013, 2016, 2020) who used evolutionary game theory to demonstrate that, in a wide range of contexts, the evolutionary stable forms of behaviour derive from either self-interested motivation or application of the Kantian categorical imperative to "act only according to that maxim through which you can at the same time will that it become a universal law". More generally, they suggest that individuals might act as imperfect Kantians, using an objective function which is a weighted average of the objective functions underlying the two stable forms of behaviour. In this paper we have explored the implications of assuming that countries act as imperfect Kantians in taking their decisions on emissions and membership.

Another motivation for this paper is the empirical observation that a growing number of people, notably young people, deliberately reduce their carbon footprint in an effort to do the morally right thing with regard to the imminent serious world-wide climate damage although they know that their emission reduction has hardly any effect on global emissions and reduces their non-Kantian utility. In their role as voters, they call on their governments to be serious about the reduction of domestic emissions and to play an active role towards an effective international climate agreement. Against this background, our paper also addresses the issue of whether trying to tackle problems such as climate change is best handled through government-level actions rather than persuading individuals, especially consumers, to make their choices in a

more moral form. We posed this as a question whether seeking to encourage individuals to act more morally is a substitute or complement to government-level actions to join an IEA.

Our key findings are that when countries act as imperfect Kantians with respect only to emissions, the resulting IEA game is iso-morphic to the conventional model in which countries act in a self-interested fashion, with the same pessimistic conclusion about IEA membership. When countries also act as imperfect Kantians with respect only to membership, we show that it is always possible to achieve the grand coalition when the weight given to acting in a Kantian fashion never exceeds 0.5 and, for some cases, could be less than 25%. The important implication of this result is that trying to form IEAs and trying to encourage individuals and their governments to act more morally are complementary approaches to trying to achieve the first-best outcome, not substitutes.

There are a number of important areas for future research, and we note three. The first is that we have assumed that all countries are identical, and it would obviously be important to explore the implications of what would happen when countries differ some respect but seek to act in a Kantian fashion. An obvious source of difference between countries is their size, their benefit functions and damage cost functions. The assumption made in this paper that perfect Kantians do the same thing may not make sense when countries differ in such respects. Van Long (2021) cites the relevant literature and applies that to his study of dynamic models of exploitation of a renewable resource, and it might be appropriate to apply his approach to our two-stage game model of IEA formation. A further aspect of enriching the model in this way would be to recognise that damage costs experienced by a country can be moderated

A second extension is that we treat countries as a single entity, whereas emissions are the results of decisions by a large number of organisations (households, producers, retailers etc) influenced, to different extents, by government policies. This may matter less given our assumption that countries are identical, but becomes important when countries differ in various respects. One important difference concerns the degree of democracy, where more autocratic can enforce policies in ways not available to more democratic governments.

A final extension, motivated by the argument in de Zeeuw (2008) that it is important to study the implications of adopting a more cooperative model of government behaviour using a dynamic model of environmental damages, would be to study the implications of imperfect Kantian behaviour when damage costs depend on the stock of greenhouse gases, not the flow of such gases.

References

Alger, J. and J.W. Weibull (2020): Morality: evolutionary foundations and policy implications, in Basu,K., Rosenblatt, D. and C. Sepulveda (eds.), *The State of Economics, the State of the World*, MIT Press.

Alger, J. and J.W. Weibull (2016): Evolution and Kantian morality, *Games and Economic Behavior* 98, 56-67.

Alger, J. and J.W. Weibull (2013): Homo moralis - preference evolution under incomplete information and assortative matching, *Econometrica* 81, 2269-2302.

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

Bernauer, T., Gampfer, R., Meng, T. and Y.-S. Su (2016): Could more civil society involvement increase public support for climate policy-making? Evidence from a survey experiment in China, *Global Environmental Change*, 40, 1-12.

Buchholz, W., Peters, W. and A. Ufert (2018): International environmental agreements on climate protection: A binary choice model with heterogeneous agents, *Journal of Economic Behavior and Organization* 154, 191-205.

Carraro, C. and D. Siniscalco (1993): Strategies for the international protection of the environment, *Journal of Public Economics* 52, 309-328.

Carraro, C. and D. Siniscalco (1991): Strategies for the international protection of the environment, CEPR Discussion Paper 568.

Chander, P. and H. Tulkens (1997): The core of an economy with multilateral environmental externalities, *International Journal of Game Theory*, 26, 379-401.

Dasgupta, P., D. Southerton, A. Ulph, and D. Ulph (2016): Consumer behaviour with environmental and social externalities, *Environmental and Resource Economics*, 65, 191-226.

Daube, M. and D. Ulph (2016): Moral behaviour, altruism and environmental policy, *Environmental and Resource Economics* 63, 505-522.

De Cara, S. and G. Rotillon (2001): Multi greenhouse gas international agreements, mimeo.

De Zeeuw, A. (2008): Dynamic effects on the stability of international environmental agreements, *Journal of Environmental Economics and Management*, 55, 163-174.

Diamantoudi, E. and E. Sartzetakis (2006): Stable international environmental agreements: An analytical approach, *Journal of Public Economic Theory* 8, 247-263.

Diamantoudi, E. and E. Sartzetakis (2015): International environmental agreements: coordinated action under foresight, *Economic Theory*, 59, 527-546.

Diamantoudi, E. and E. Sartzetakis (2018): International environmental agreements: the role of foresight, *Environmental and Resource Economics*, 71, 241-257.

Eichner T. and R. Pethig (2022): International environmental agreements when countries behave morally, CESifo Working Paper No 10090.

Eichner, T. and R. Pethig (2021): Climate policy and moral consumers, *Scandinavian Journal of Economics* 123, 1190-1226.

Eichner, T. and R. Pethig (2015): Is trade liberalization conducive to the formation of climate coalitions?, *International Tax and Public Finance* 22, 932-955.

Finus, M. (2003): Stability and design of international environmental agreements: The Case of transboundary pollution, in Folmer, H. and T. Tietenberg (eds.): *International Yearbook of Environmental and Resource Economics*, 2003/4, Edward Elgar, 82-158.

Finus, M. and A. Caparros (2015): *Handbook on Game Theory and International Environmental Cooperation: Essential Readings*, Edward Elgar.

Finus, M., Furini, F. and A. Rohrer (2021a): The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership, *Journal of Environmental Economics and Management* 109, 102461.

Finus, M., Furini, F. and A. Rohrer (2021b): International environmental agreements and the paradox of cooperation: Revisiting and generalising some previous results, *GEP 2021-5*, University of Graz.

Finus, M. and S. Maus (2008): Modesty may pay!, *Journal of Public Economic Theory* 10, 801-826.

Finus, M. and B. Rundshagen (2001): Endogenous coalition formation in global pollution control, working paper No. 43, 2001, Milan: Fondazione Eni Enrico Mattei.

Grafton, R.Q., Kompas, T. and N. van Long (2017): A brave new world? Kantian-Nashian interaction and the dynamics of global climate change mitigation, *European Economic Review* 99, 31-42.

Gaus, G. (2010): *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, Cambridge University Press.

Johnson, R. and A. Cureton (2021): Kant's Moral Philosophy, in Zalta, E. (ed.): *The Stanford Encyclopedia of Philosophy.*

Hoel, M. (1992): International environmental conventions: The case of uniform reductions of emissions, *Environmental and Resource Economics* 2, 141-159.

Kamarack, E. (2019): The challenging politics of climate change, Brookings Institute Report.
Kant, I. (1785): Grundlegung zur Metaphysik der Sitten. [In English: Groundwork of the Metaphysics of Morals. 1964. New York: Harper Torch books.]

Laffont, J.-J. (1975): Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics, *Economica* 42, 430-437.

Lange, A. and C. Vogt (2003): Cooperation in international environmental negotiations due to a preference for equity, *Journal of Public Economics* 87, 2049-2067.

Nawrotzki, R.J. (2012): The politics of environmental concern - a cross-country analysis, *Organisation and Environment* 25, 286-301.

Nkuyia, B. (2020): Stability of international environmental agreements under iso-elastic utility, *Resource and Energy Economics* 59, 101-128.

Nowakowski, A. and A.J. Oswald (2020): Do Europeans care about climate change? An illustration of the importance of data on human feelings, *IZA Discussion Paper* 13660.

Nyborg, K. (2018a): Social norms and the environment, *Annual Review of Resource Economics* 10, 405-423.

Nyborg, K. (2018b): Reciprocal climate negotiators, *Journal of Environmental Economics and Management* 92, 707-725.

Roemer, J.E. (2015): Kantian optimization. A microfoundation for cooperation, *Journal of Public Economics* 127, 45-57.

Roemer, J.E. (2010): Kantian equilibrium, *Scandinavian Journal of Economics* 112, 1-24.

Rogna, M. and C. Vogt (2020): Coalition formation with optimal transfers when players are heterogenous and inequality averse, *Ruhr Economic Papers* 865.

Romeijn, J. (2020): Do political parties listen to the(ir) public? Public opinion - party linkage on specific policy issues, *Party Politics* 26, 426-436.

Rubio, S.J. and A. Ulph (2006): Self-enforcing agreements and international trade in greenhouse emission rights, *Oxford Economic Papers* 58, 233-263.

Rubio, S. and B. Casino (2002): A note on cooperative versus non-cooperative strategies in international pollution control, *Resource and Energy Economics* 24, 251-261.

Sudbury-Riley, L. and F. Kohlbacher (2016): Ethically minded consumer behavior: scale review, development and validation, *Journal of Business Research* 69, 2697-2710.

UNEP (United Nations Environment Programme) (2019): The emissions gap report 2019.

Van Long N. (2021): Dynamic games of common-property resource exploitation when self-image matters. In: Dawid H., Arifovic J. (eds.), *Dynamic Analysis in Complex Economic Environments. Dynamic Modeling and Econometrics in Economics and Finance* 26. Springer, Cham.

Van Long, N. (2020): A dynamic game with interaction between Kantian players and Nashian players, in Pineau, P.-O., Sigué, S. and S. Taboubi (eds.), *Games in Management Science: Essays in Honor of Georges Zaccour*, Springer:, Switzerland.

Van Long, N. (2016): The impacts of the other-regarding preferences and ethical choice on environmental outcomes: A review of the literature, Scientific Series 2016s-10, CIRANO, Montreal.

Van der Pol, T., Weikard, H.-P. and E. van Ireland (2012): Can altruism stabilize international climate agreements, *Ecological Economics* 81, 33-59.

Vogt, C. (2016): Climate coalition formation when players are heterogeneous and inequality averse, *Environmental and Resource Economics* 65, 33-39.
White, K., Habib, R. and D. Hardisty (2019): How to shift consumer behaviors to be more sustainable: a literature review and guiding framework,

Appendix A: Equilibrium Outputs of Model with Quadratic Benefit Function and Quadratic Damage Cost Function.

From the definition of *B*(.) and *D*(.) it is straightforward to derive the following.

First Best, Fully Cooperative and Self-Interested Non-Cooperative Equilibria

$$e^{SO} = e^{FC} = \frac{\beta}{(1+\delta n^2)};\ W^{SO} = W^{FC} = \frac{0.5\beta^2}{(1+\delta n^2)^2}$$

$$e^{NC} = \frac{\beta}{(1+\delta n)} > e^{SO};\ \ W^{NC} = \frac{0.5\beta^2(1+2\delta n-\delta n^2)}{(1+\delta n)^2} < W^{SO}$$

Imperfect Kantian Non-Cooperative Equilibrium

$\hat{e}(\kappa) = \frac{\beta}{1+\delta[\kappa n^2+(1-\kappa)n]}$. It is clear that $\frac{\partial \hat{e}(.)}{\partial \kappa} < 0$, falling from $\hat{e}(0) = e^{NC}$ to $\hat{e}(1) = e^{*}$.

We do not give an explicit expression for welfare as it is rather messy.

IEAs with Kantian Behaviour – Stage 2 Equilibrium Emissions

For $m = 1$, all countries are in the fringe, so $\tilde{e}^{f}(1,\kappa_{\varepsilon}) = \hat{e}(\kappa_{\varepsilon})$; for $m = n$, all countries are in the (grand) coalition, so $\tilde{e}^{c}(n,\kappa_{\varepsilon}) = e^{SO}$.

For $2 \leq m \leq n-1$, the first-order conditions are:

$$\beta - [1 + \delta\kappa_{\varepsilon}n^2 + (1-\kappa_{\varepsilon})\delta m^2]e^{c} - [(1-\kappa_{\varepsilon})\delta m(n-m)]e^{f} = 0 \qquad \text{(A1a)}$$

$$\beta - [1 + \delta\kappa_{\varepsilon}n^2 + (1-\kappa_{\varepsilon})\delta(n-m)]e^{f} - (1-\kappa_{\varepsilon})\delta m e^{c} = 0 \qquad \text{(A1b)}$$

Define:

$$\phi(\kappa_{\varepsilon}) = \frac{1+\delta\kappa_{\varepsilon}n^2}{(1-\kappa_{\varepsilon})\delta};\ \phi'(\kappa_{\varepsilon}) = \frac{1+\delta n^2}{\delta(1-\kappa_{\varepsilon})^2} > 0;\ \phi(0) = \frac{1}{\delta} > 0.25(n-1)^2;\ \kappa_{\varepsilon} \to 1 \Longrightarrow \phi(.) \to \infty$$

Solving (A1a,b) yields:

$$\tilde{e}^{c}(m,\kappa_{\varepsilon}) = \frac{\beta[\phi(\kappa_{\varepsilon})-(n-m)(m-1)]}{(1+\delta\kappa_{\varepsilon}n^2)[\phi(\kappa_{\varepsilon})+(n+m^2-m)]} > 0; \qquad \text{(A2a)}$$

$$\tilde{e}^{f}(m,\kappa_{\varepsilon}) = \frac{\beta[\phi(\kappa_{\varepsilon})+m^2-m]}{(1+\delta\kappa_{\varepsilon}n^2)[\phi(\kappa_{\varepsilon})+(n+m^2-m)]} > 0; \qquad \text{(A2b)}$$

$$\tilde{e}^{a}(m,\kappa_{\varepsilon}) = [m * \tilde{e}^{c}(m,\kappa_{\varepsilon}) + (n-m)\tilde{e}^{f}(m,\kappa_{\varepsilon})]/n$$

$$= \frac{\beta n\phi(\kappa_{\varepsilon})}{(1+\delta\kappa_{\varepsilon}n^2)[\phi(\kappa_{\varepsilon})+(n+m^2-m)]} \qquad \text{(A2c)}$$

It is straightforward to see that $\tilde{e}^{f}(m,\kappa_{\varepsilon}) > \tilde{e}^{c}(m,\kappa_{\varepsilon}),\ 2 \leq m \leq n-1$.

We now determine the signs of $\frac{\partial \tilde{e}^{i}(.)}{\partial x},\ i = c,f,a; x = m,\kappa_{\varepsilon}$. We ignore the parameter β which is a simple scaling factor. To save notation, we will also ignore the superscript ~ on equilibrium emissions, and the subscript ε on κ_{ε}.

Variations with respect to *m*.

It is straightforward to see that $\frac{\partial e^f(.)}{\partial m} > 0$; $\frac{\partial e^a(.)}{\partial m} < 0 \quad \forall\, 2 \le m \le n-1; 0 < \kappa < 1.$

$$\frac{\partial e^c}{\partial m} = -\frac{[\phi(\kappa)+n+m^2-m](n+1-2m)+[\phi(\kappa)-(n-m)(m-1)](2m-1)}{[\phi(\kappa)+n+m^2-m]^2} \qquad \text{(A3)}$$

The sign of $\frac{\partial e^c}{\partial m}$ depends on the sign of the numerator in (A3), which we denote by $\chi(m,\kappa)$. After some simplification:

$$\chi(m,\kappa) = n[\phi(\kappa) + n - m^2] \qquad \text{(A4)}$$

Define: $\bar{\kappa}\ s.t.\ \phi(\bar{\kappa}) + n - (n-1)^2 = 0$; after some rearrangement we have:

$$\bar{\kappa} = \frac{(n-1)^2\left(1-\frac{\bar{\delta}}{4\delta}\right)-n}{n^2+(n-1)^2-n} \qquad \text{(A5)}$$

For $0 < \kappa < \bar{\kappa}$, from (A4) define $\hat{m}(\kappa)$ as the smallest integer greater than or equal to $\sqrt{\phi(\kappa)+n}$, the value of *m* as a real variable that solves $\chi(m,\kappa) = 0$, and hence the value of *m* at which $\frac{\partial e^c}{\partial m}$ reaches a minimum.

Finally, we ask if there is a range of values of *m* for which $e^c(m,\kappa) \le e^{SO}$, which clearly must include $\hat{m}(\kappa)$. So we ask: for any given $\kappa, 0 < \kappa < \bar{\kappa}$, for what values of *m* does $e^c(m,\kappa) = e^{SO}$? Thus we require:

$$\frac{\phi(\kappa)-(n-m)(m-1)}{(1+\delta\kappa n^2)[\phi(\kappa)+n+m^2-m)]} = \frac{1}{1+\delta n^2}$$

i.e. $\quad m^2 - \nu m + \mu = 0 \qquad \text{(A6i)}$

where: $\nu = 1 + \frac{1+\delta n^2}{\delta(1-\kappa)n};\quad \mu = \phi(\kappa) + n) \qquad \text{(A6ii)}$

Thus $e^c(m,\kappa) \le e^{SO} \Leftrightarrow \underline{m}(\kappa) \le m \le \overline{m}(\kappa)$ where:

$$\underline{m}(\kappa) = 0.5\left[\nu - \sqrt{\nu^2 - 4\mu}\right]; \qquad \text{(A7i)}$$

$$\overline{m}(\kappa) = 0.5\left[\nu + \sqrt{\nu^2 - 4\mu}\right]; \text{ and} \qquad \text{(A7ii)}$$

Now it is straightforward to check that $m = n$ solves (A6i) for all κ so we can set $\overline{m}(\kappa) = n$.

We summarise the results for the effects of variations in *m* on equilibrium Stage 2 emissions.

Result A1

Varying membership affects Stage 2 equilibrium emissions as follows:

(i) $\quad \frac{\partial \tilde{e}^a}{\partial m} < 0 \quad \forall m, \kappa$

(ii) $\quad \frac{\partial \tilde{e}^f}{\partial m} > 0 \quad \forall m, \kappa$

(iii) $\quad For\ 1 > \kappa \ge \overline{\kappa}, \frac{\partial \tilde{e}^c}{\partial m} < 0,\ \forall m,\ 2 \le m \le n-1$

(iv) $\quad For\ 0 < \kappa < \overline{\kappa},\ \frac{\partial \tilde{e}^c}{\partial m} \le 0 \Leftrightarrow 2 \le m < \hat{m}(\kappa);\ \frac{\partial \tilde{e}^c}{\partial m} > 0 \Leftrightarrow \hat{m}(\kappa) \le m \le n.$

(v) $\quad For\ 0 < \kappa < \overline{\kappa},\ \tilde{e}^c(m,\kappa) \le e^{SO} \Leftrightarrow \underline{m}(\kappa) \le m \le n$

Variations with respect to κ .

We consider first $\frac{\partial e^a}{\partial \kappa}$; from (A2c) we have:

$$\frac{\partial e^a}{\partial \kappa} = \frac{\phi(\kappa)+n+m^2-m-(1-\kappa)\phi'}{[(1-\kappa)(\phi(\kappa)+n+m^2-m]^2} \qquad \text{(A8)}$$

The sign of $\frac{\partial e^a}{\partial \kappa}$ depends on the sign of the numerator in (A8) which we denote $\psi(m,\kappa)$ where

$$\psi(m,\kappa) = \frac{(1+\delta\kappa n^2)}{\delta(1-\kappa)} + n + m^2 - m - \frac{(1-\kappa)(1+\delta n^2)}{\delta(1-\kappa)^2}$$

$$= n + m^2 - m - n^2 = m(m-1) - n(n-1)$$

Hence: $\forall\kappa \quad \frac{\partial e^a}{\partial \kappa} \leq 0 \quad \Leftrightarrow m \leq n.$

We now consider $\frac{\partial e^f}{\partial \kappa}$. From (A2b) we have:

$$\frac{\partial e^f}{\partial \kappa} = \frac{n(1+\delta\kappa n^2)\phi' - \delta n^2[\phi(\kappa)+m^2-m][\phi(\kappa)+n+m^2-m]}{[(1+\delta\kappa n^2)(\phi(\kappa)+n+m^2-m)]^2} \qquad \text{(A9a)}$$

The sign of $\frac{\partial e^f}{\partial \kappa}$ depends on the sign of the numerator in (A9a), which, ignoring a common factor, $n\delta$, we denote by $\omega(m,\kappa)$ where:

$$\omega(m,\kappa) = \frac{(1+\delta\kappa n^2)(1+\delta n^2)}{\delta^2(1-\kappa)^2} - n[\phi(\kappa) + m(m-1)][\phi(\kappa) + n + m(m-1)] \quad \text{(A9b)}$$

Note that $1 + \delta n^2 = 1 + \delta\kappa n^2 + \delta n^2(1-\kappa)$

Then (A9b) becomes:

$$\omega(m,\kappa) = \phi(\kappa)[\phi(\kappa) + n^2]$$

$$-n[\phi(\kappa)^2 + 2(m^2-m)\phi(\kappa) + n\phi(\kappa) + (m^2-m)(n+m^2-m)] \qquad \text{(A9c)}$$

$$\omega(m,\kappa) = -(n-1)\phi(\kappa)^2 - 2n(m^2-m)\phi(\kappa) - n(m^2-m)(n+m^2-m) \quad \text{(A9d)}$$

Hence:

$$\frac{\partial e^f}{\partial \kappa} < 0 \quad \forall m,\kappa$$

Finally, we consider $\frac{\partial e^c}{\partial \kappa}$. To save notation, denote $\lambda = \lambda(m) = (n-m)(m-1) < mn$. Then:

$$\frac{\partial e^c}{\partial \kappa} = \frac{mn(1+\delta\kappa n^2)\phi' - \delta n^2[\phi(\kappa)-\lambda][\phi(\kappa)-\lambda+mn)]}{[(1+\delta\kappa n^2)(\phi(\kappa)+n+m^2-m)]^2} \qquad \text{(A10a)}$$

The sign of $\frac{\partial e^c}{\partial \kappa}$ depends on the sign of the numerator in (A10a), which, ignoring a common factor, $n\delta$, we denote by $\varphi(m,\kappa)$ where:

$$\varphi(m,\kappa) = m\frac{(1+\delta\kappa n^2)(1+\delta n^2)}{\delta^2(1-\kappa)^2} n[\phi(\kappa) - \lambda][\phi(\kappa) + (mn-\lambda)] \qquad \text{(A10b)}$$

As in (A9b), (A10b) becomes

$$\varphi(m,\kappa) = m\phi(\kappa)(\phi(\kappa)+n^2) -$$
$$n[\phi(\kappa)^2 - \phi(\kappa)(2\lambda - mn) + \lambda(\lambda - mn)]$$
$$\varphi(m,\kappa) = -(n-m)\phi(\kappa)^2 + 2n\lambda\phi(\kappa) + n\lambda(n + m^2 - m) \qquad \text{(A10c)}$$

Substituting back for λ and ignoring a common term, $(n - m)$ we get

$$\varphi(m,\kappa) = -[\phi(\kappa)]^2 + 2\phi(\kappa)n(m-1) + n(m-1)(n + m^2 - m) \qquad \text{(A10d)}$$

The sign of $\varphi(m,\kappa)$ is ambiguous. We know that $\kappa \longrightarrow 1 \Rightarrow \phi(\kappa) \longrightarrow \infty \Rightarrow \varphi(m,\kappa) < 0\ \forall m$. However, for lower values of κ, and for relatively large values of m, $\varphi(m,\kappa)$ could be positive. To investigate further, define $\check{\kappa}(m)$ as the value of κ for which $\varphi(m,\kappa) = 0$. We determine this in two stages. First, we solve (A10d) for the value of $\check{\phi}(m)$ for which $\varphi(m,\kappa) = 0$. So:

$$\check{\phi}(m) = n(m-1) + \sqrt{[n(m-1)]^2 + [n(m-1)][n + m^2 - m]}. \qquad \text{(A10e)}$$

We now solve for value of $\check{\kappa}(m)$ for which $\phi[\check{\kappa}(m)] = \check{\phi}(m)$. Hence:

$$\check{\kappa}(m) = \frac{\check{\phi}(m) - 1/\delta}{\check{\phi}(m) + n^2} \qquad \text{(A10f)}$$

It is straightforward to see that $\check{\phi}'(.) > 0,\ \check{\kappa}'(m) > 0$. (A10f) confirms that if m, and hence $\check{\phi}(m)$, is sufficiently small then $\check{\kappa}(m)$ can be negative, in which case $\frac{\partial e^c}{\partial \kappa} < 0 \quad \forall \kappa \geq 0$. For larger values of m, $\check{\kappa}(m) > 0$, in which case $\frac{\partial e^c}{\partial \kappa} > 0, 0 \leq \kappa < \check{\kappa}(m)$, and $\frac{\partial e^c}{\partial \kappa} \leq 0, \check{\kappa}(m) \leq \kappa \leq 1$.

We summarise the results for the effects of variations in κ on equilibrium Stage 2 emissions as follows.

Result A2

(i) $\frac{\partial e^a}{\partial \kappa} \leq 0 \Leftrightarrow m \leq n,\ \forall \kappa$

(ii) $\frac{\partial e^f}{\partial \kappa} < 0 \quad \forall m, \kappa$

(iii) If $\check{\kappa}(m) \leq 0, \frac{\partial e^c}{\partial \kappa} < 0\ \forall \kappa$; if $\check{\kappa}(m) > 0, \frac{\partial e^c}{\partial \kappa} \leq 0 \Leftrightarrow 1 \geq \kappa \geq \check{\kappa}(m)$

This completes the proofs of Results 4(a) and 4(b) in Section 3.1.1.

ONLINE APPENDIX B

RESULTS USING QUADRATIC BENEFIT FUNCTION, AND LINEAR DAMAGE COST FUNCTION

In this Online Appendix we present results for our model using a linear damage cost function rather than the quadratic damage cost function used in Section 3 of our paper. This is the same damage cost function used by Eichner and Pethig (2022) to derive their numerical results. We assume that $B(e) = \beta e - 0.5e^2$ and $D[\sum_{j\in N} e_j] = \delta[\sum_{j\in N} e_j]$. To ensure we always have non-negative emissions, we assume that $0 < \delta < \beta/n$. The key results for the first-best, non-cooperative self-interested equilibrium, the non-cooperative equilibrium with Kantian behaviour, and our model of IEAs with Kantian behaviour towards emissions and membership are derived in Appendix C. As we note there, all the key outputs, $\hat{\tilde{e}}(\kappa), \hat{\widetilde{W}}(\kappa)$, $\tilde{\tilde{e}}^j(\kappa_\varepsilon, \kappa_\mu), \tilde{\widetilde{W}}^j(\kappa_\varepsilon, \kappa_\mu)\ j = c, f, a$ do not depend on the parameters β, δ.

We begin by presenting the equilibrium outputs for the non-cooperative equilibrium with Kantian behaviour, namely $\hat{\tilde{e}}(\kappa) = \kappa,\ \ \hat{\widetilde{W}}(\kappa) = 1 - (1-\kappa)^2$. The proportion of the emissions gap closed increases linearly in κ, while the proportion of the welfare gap closed increases at a decreasing rate as κ increases.

We now turn to address the four questions discussed in Section 3 of our paper.

Question (1): Behaviour of Stage 2 Equilibrium Emissions and Welfare.

From Appendix C we have:

$\tilde{e}^c(m, \kappa_\varepsilon) = e^{SO} + \delta(1-\kappa_\varepsilon)(n-m);\ \tilde{e}^f(m, \kappa_\varepsilon) = e^{SO} + \delta(1-\kappa_\varepsilon)(n-1)$

$\tilde{e}^a(m, \kappa) = e^{SO} + \left(\frac{\delta}{n}\right)[(1-\kappa)(n-m)(n+m-1)];$

$\widetilde{W}^c(m, \kappa) = W^{so} - 0.5\delta^2(1-\kappa)[(1-\kappa)(n-m)^2 + 2(n-m)(m-1)];$

$\widetilde{W}^f(m, \kappa) = W^{so} - 0.5\delta^2(1-\kappa)[(1-\kappa)(n-1)^2 - 2m(m-1)];$

$$\widetilde{W}^a(m, \kappa) = W^{so} - 0.5\delta^2(1-\kappa)^2(n-m)\left[\frac{m(n-m) + (n-1)^2}{n}\right];$$

where: $e^{SO} = \beta - \delta n;\ \ W^{SO} = 0.5(e^{SO})^2$

Then it is straightforward to show that:

Result B1

The effects of variations in membership (m) *and Kantian weight on emissions* (κ_ε) *on Stage 2 equilibrium and welfare are as follows:*

$\frac{\partial \tilde{e}^c(m,\kappa)}{\partial \kappa} = -\delta(n-m) < 0;\ \ \frac{\partial \tilde{e}^c(m,\kappa)}{\partial m} = -\delta(1-\kappa) < 0;$ B1(i)

$\frac{\partial \tilde{e}^f(m,\kappa)}{\partial \kappa} = -\delta(n-1) < 0;\ \frac{\partial \tilde{e}^f(m,\kappa)}{\partial m} = 0;$ B1(ii)

$$\frac{\partial \tilde{e}^a(m,\kappa)}{\partial \kappa} = -\frac{\delta}{n}(n-m)(n+m-1) < 0; \qquad \text{B1(iii)}$$

$$\frac{\partial \tilde{e}^a(m,\kappa)}{\partial m} = -\frac{\delta}{n}(1-\kappa)(1-2m\kappa) < 0 \qquad \text{B1(iv)}$$

$$\frac{\partial \widetilde{W}^c(m,\kappa)}{\partial \kappa} = \delta^2[(n-m)(1-\kappa)+(m-1)] > 0; \qquad \text{B1(v)}$$

$$\frac{\partial \widetilde{W}^c(m,\kappa)}{\partial m} = 0.5\delta^2(1-\kappa)[(n-1)-2\kappa(n-m)]; \qquad \text{B1(vi)}$$

$$\frac{\partial \widetilde{W}^f(m,\kappa)}{\partial \kappa} = (\frac{\delta}{n-1})^2\left[1-\kappa-\frac{m(m-1)}{(n-1)^2}\right] \qquad \text{B1(vii)}$$

$$\frac{\partial \widetilde{W}^f(m,\kappa)}{\partial m} = \delta^2[(1-\kappa)(2m-1)] > 0; \qquad \text{B1(viii)}$$

$$\frac{\partial \widetilde{W}^a(m,\kappa)}{\partial \kappa} = \delta^2(1-\kappa)(n-m)\left[\frac{m(n-m)+(n-1)^2}{n}\right] > 0; \qquad \text{B1(ix)}$$

$$\frac{\partial \widetilde{W}^a(m,\kappa)}{\partial m} = \frac{\delta^2}{2n}(1-\kappa)^2[2m(n-m)+(n-1)^2-(n-m)^2] > 0;$$

Emissions fall steadily from non-cooperative to fully cooperative level as κ rises for coalition and fringe countries and hence for an average country. Emissions for a coalition country fall as membership increases, but, for a fringe country, emissions are independent of membership size; obviously average emissions fall as membership rises.

Welfare is more interesting. For an average country welfare rises with both κ and m. For a coalition country, welfare rises with Kantian weight, but it increases with membership iff $\kappa(n-m) < 0.5(n-1)$; the intuition is that because fringe countries keep emissions constant, as membership increases the gap between coalition and fringe country emissions widens, i.e. free riding increases, so this hurts coalition countries. Consequently, welfare of a fringe country rises as membership increases. But, when κ increases, welfare of a fringe country falls when κ is larger than a critical level $\underline{\kappa}(m,n)$ where:

$$\underline{\kappa}(m,n) = 1-\frac{m(m-1)}{(n-1)^2}.$$

For large *n* and small *m*, then in the imperfect Kantian emissions only model, $\underline{\kappa}(m,n)$ will be quite close to 1. The intuition is related to the fact that fringe countries are free riding the emission reductions of coalition countries, and can get welfare which exceeds first-best when *m* is high.

Question (2): Size of Equilibrium IEA.

The second question is what is the size of the equilibrium IEA, and for what parameter values is the grand coalition an equilibrium. In Appendix C we show that the stability function is:

$$\sigma(m,\kappa_\varepsilon,\kappa_\mu) = 0.5\delta^2(1-\kappa_\varepsilon)^2[\kappa_\mu(n-1)^2-(1-\kappa_\mu)(m-1)(m-3)] \qquad \text{(B1)}$$

It is straightforward to confirm the following special cases:

(a) Non-Kantian: $\sigma(m,0,0) = (m-1)(3-m) \Longrightarrow \tilde{m}(0,0) = 3$;

(b) Perfect Kantian wrt emissions only: $\sigma(m,1,\kappa_\mu) = 0 \Longrightarrow \tilde{m}(1,\kappa_\mu) = n$;

(c) Perfect Kantian wrt membership only: $\sigma(m, \kappa_\varepsilon, 1) = (1-\kappa_\varepsilon)^2(n-1)^2 \Longrightarrow \tilde{m}(\kappa_\varepsilon, 1) = n$.

(d) Perfect Kantian wrt emissions and membership: $\sigma(m, 1, \kappa_\mu) = 0 \Longrightarrow \tilde{m}(1, \kappa_\mu) = n$;

For the general case, $0 < \kappa_\varepsilon, \kappa_\mu < 1$, it is clear from (B1) that (i) $\sigma(.) > 0, m \leq 3$; (ii) $\frac{\partial \sigma(.)}{\partial m} < 0, m > 2$, so there is a unique stable IEA no less than 3; (iii) the size of the stable IEA, defined as $\tilde{m}(\kappa_\mu)$, depends only on κ_μ. We show in Appendix C that $\frac{d\tilde{m}(\kappa_\mu)}{d\kappa_\mu} > 0$ and that $\tilde{m}(0.5) = n$, so the grand coalition is the stable equilibrium when κ_μ is at least 0.5. To get further results it is useful to define $\underline{\kappa}_\mu(m)$ as the minimum value of κ_μ required to stabilise an IEA of size *m* out of *n* countries, *m* = 4,…,*n*. Then, from (B1),

$$\underline{\kappa}_\mu(m) = \frac{(m-1)(m-3)}{(n-1)^2+(m-1)(m-3)}. \qquad \text{(B2)}$$

It is clear that $\underline{\kappa}_\mu(m)$ is increasing in *m*, so, not surprisingly, a higher Kantian weight on membership is needed to stabilise a larger coalition. In the limit, the minimum Kantian weight needed to make the grand coalition stable is:

$$\underline{\kappa}_\mu(n) = \frac{(n-3)}{2(n-2)} < 0.5 \qquad \text{(B3)}$$

Hence:

Result B2

With quadratic benefit function and linear damage cost function:

(i) the size of the equilibrium IEA depends only on the Kantian weight on membership, κ_μ, and is an increasing function of that weight;

(ii) the Kantian weight on emissions needed to secure the grand coalition increases from 0.25 *as the number of countries increases from 4, but never exceeds* 0.5.

Thus, we have the important result that the first-best outcome can be achieved with a Kantian weight on membership that never exceeds 0.5.

Question 3: Extent to which IEAs with Imperfect Kantian Behaviour Close the Emissions and Welfare Gaps of Coalition and Fringe Countries.

Looking first at emissions gaps for coalition and fringe countries, Appendix C shows that:

$$\tilde{e}^c(\kappa_\varepsilon, \kappa_\mu) = 1 - \frac{(1-\kappa_\varepsilon)\left(n-\tilde{m}(\kappa_\mu)\right)}{(n-1)}; \quad \frac{\partial \tilde{e}^c}{\partial \kappa_\varepsilon} > 0; \quad \frac{\partial \tilde{e}^c}{\partial \kappa_\mu} > 0; \quad \tilde{e}^f(\kappa_\varepsilon) = \kappa_\varepsilon; \quad \frac{\partial \tilde{e}^f}{\partial \kappa_\varepsilon} > 0$$

Thus, the emissions gap for fringe countries is constant, and equal to that which would arise in the Kantian non-cooperative equilibrium when $\kappa = \kappa_\varepsilon$. If $\tilde{m}(\kappa_\mu) = 1$, the emissions gap for coalition countries would also be κ_ε, and as κ_μ rises the coalition's emissions gap shrinks, with $\tilde{e}^c(.)$ reaching 1 when $\tilde{m}(\kappa_\mu) = n$, so the grand coalition is attained and first-best emissions is attained.

Turning to the welfare gaps for coalition and fringe countries, Appendix C presents the formulae for $\tilde{\tilde{W}}^c(\kappa_\varepsilon, \kappa_\mu)$, $\tilde{\tilde{W}}^f(\kappa_\varepsilon, \kappa_\mu)$. We show that $\frac{\partial \tilde{\tilde{W}}^j(.)}{\partial \kappa_\varepsilon} > 0$, j = *c,f*. Beginning with a fringe country, if $\tilde{m}(\kappa_\mu) = 1$, its welfare gap equals the level that would arise in the Kantian non-cooperative equilibrium when $\kappa = \kappa_\varepsilon$, as we would expect. As κ_μ increases, the extent to which a fringe country's welfare gap shrinks increases, $\frac{\partial \tilde{\tilde{W}}^f}{\partial \kappa_\mu} > 0$. Indeed, as we will show below, it is possible that for larger values of κ_μ, but not so large that the grand coalition is an equilibrium, $\tilde{\tilde{W}}^f(\kappa_\varepsilon, \kappa_\mu)$ can exceed 1, so fringe welfare rises above the first-best level of welfare. The rationale is simply that fringe countries free-ride on the efforts of coalition countries to cut their emissions (as we have seen, fringe countries keep their emissions constant at the level in the non-cooperative Kantian equilibrium), so, as the size of the coalition increases, the free-riding benefits from a larger coalition are shared by a smaller number of fringe countries, and so can exceed first-best welfare.

Turning to coalition countries, if $\tilde{m}(\kappa_\mu) = n$, then the welfare gap for a coalition country is fully closed: $\tilde{\tilde{W}}^c(.) = 1$, as we would expect. To assess how the welfare of a coalition country varies with the Kantian weight on membership, we show, in Appendix C, that $\frac{\partial \tilde{\tilde{W}}^c}{\partial \kappa_\mu} = \chi[\tilde{m}(\kappa_\mu) - \frac{1+n\kappa_\varepsilon}{1+\kappa_\varepsilon}]$, where χ is positive. It follows that the sign of $\frac{\partial \tilde{\tilde{W}}^c}{\partial \kappa_\mu}$ is ambiguous; if κ_μ is small relative κ_ε, and hence $\tilde{m}(\kappa_\mu)$ is small relative to *n*, the expression in square brackets may be negative and so $\tilde{\tilde{W}}^c(.)$ may be initially decreasing in κ_μ.

To summarise:

Result B3

With quadratic benefit function and linear damage cost function:

- *(i) the emission gaps of coalition and fringe countries decrease as the Kantian weights $\kappa_\varepsilon, \kappa_\mu$ increase;*
- *(iii) the welfare gap of a fringe country decreases as $\kappa_\varepsilon, \kappa_\mu$ increase; for values κ_μ tending towards the value for which the grand coalition forms, the welfare gap of a fringe country may be negative, i.e. welfare of a fringe country exceeds first-best;*
- *(iv) the welfare gap of a coalition country decreases as κ_ε increases; however, for values of $\tilde{m}(\kappa_\mu) < \frac{1+n\kappa_\varepsilon}{1+\kappa_\varepsilon}$ the welfare gap of a coalition is increasing in κ_μ.*

To provide more detail on how coalition and fringe welfare vary with κ_μ, it is simplest to use some numerical illustrations. In Figure 1a we plot the extent of closure of the welfare gap of a coalition country, $\tilde{\tilde{W}}^c(\kappa_\varepsilon, \kappa_\mu)$, for 20 values of κ_μ between 0 and 0.5 for 3 different values of κ_ε = 0.1, 0.3 and 0.5[30]. As we see, when κ_ε = 0.1, coalition welfare increases for all values of κ_μ; when κ_ε = 0.3, then coalition welfare decreases between κ_μ = 0.01 and 0.05 and then rises

[30] We present the data on welfare in % terms, rather than between 0 and 1.

till κ_μ = 0.5 when the grand coalition forms; when κ_ε = 0.5, then coalition welfare decreases between κ_μ = 0.01 and 0.10, and then increases until the grand coalition forms. The rationale is that larger values of κ_ε mean all countries are making significant cuts in emissions; when κ_μ is small, so there are relatively few coalition countries, the extra burden of cutting emissions with no contribution from fringe countries, means coalition countries experience a modest decrease in equilibrium welfare; this is offset as κ_μ, and hence the equilibrium size of the

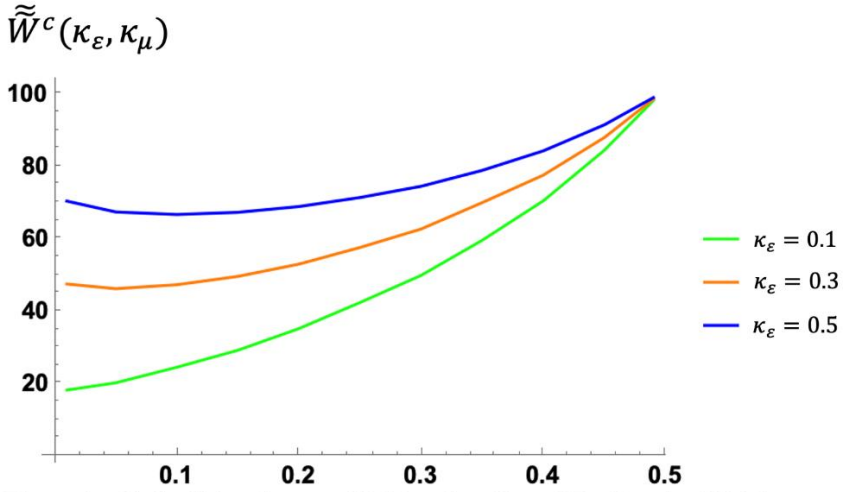

Figure 1a: % Coalition Country Welfare Gap Closed For Kantian Weights

coalition increases.

Figure 1b shows data for fringe country welfare for the same values of $\kappa_\varepsilon, \kappa_\mu$. As we know, a fringe countries welfare increases for all values of κ_μ. When κ_ε = 0.1, fringe welfare rises above 100% when κ_μ reaches 0.35 and tends towards 200% as κ_μ tends to 0.49. For values of κ_ε = 0.3 and 0.5, fringe welfare exceeds 100% at lower values of κ_μ but tends to a lower limit as κ_μ tends to 0.49.

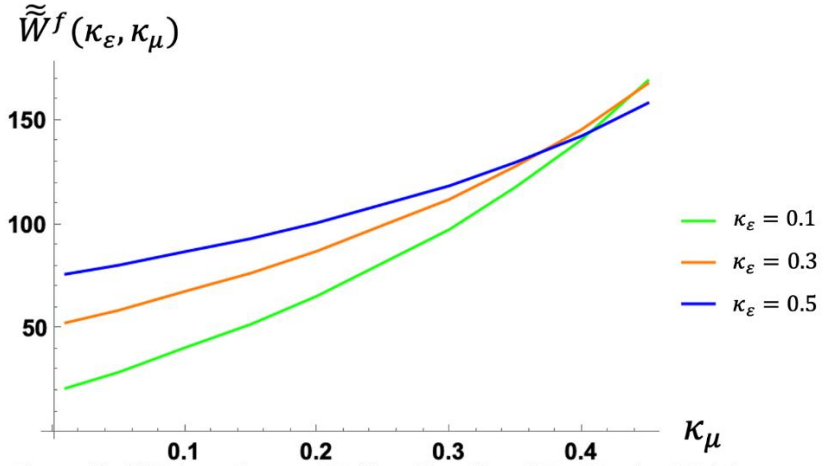

Figure 1b: % Fringe Country Welfare Gap Closed For Kantian Weights

For both coalition and fringe countries welfare increases significantly with κ_ε.

Question 4: Extent to which Overall Emissions and Welfare Gaps are closed by Countries Acting in an Imperfect Kantian Fashion and by Forming an IEA.

In this section we address the question posed by Nowakowski and Oswald (2021) about the extent to which climate change is better addressed by focusing on influencing individuals' behaviour by making them act more morally with respect to their consumption and production decisions, rather than trying to form global coalitions of countries. To analyse this question, we assume that individual moral attitudes are captured by κ_ε, while government moral attitudes are captured by κ_μ. In the absence of any attempts by governments to form IEAs, the outcome in terms of emissions and welfare is captured by an average country's emissions and welfare in a Kantian non-cooperative equilibrium. It is simplest to address this question numerically, using 3 values for κ_ε and κ_μ: 0.1, 0.3, 0.5. Table B1a presents the data for emissions.

As we see, if $\kappa_\varepsilon = \kappa_\mu$, then changing individual behaviour and negotiating an IEA make roughly equal contributions to closing the gap between the self-interested non-cooperative outcome and first best. However, if $\kappa_\mu < \kappa_\varepsilon$, then, not surprisingly trying to form an IEA, contributes significantly less than changing individual moral attitudes. On the other hand, if $\kappa_\mu > \kappa_\varepsilon$ the reverse is true; in the extreme case where $\kappa_\mu = 0.5$, $\kappa_\varepsilon = 0.1$, trying to form an IEA increases the reduction in emissions from 10% to 100%.

κ_ε	Imperfect Kantian Non-Coop Equilibrium	IEAs- κ_μ		
		0.1	0.3	0.5
0.1	10.00	20.82	49.00	100.00
0.3	30.00	38.41	60.33	100.00
0.5	50.55	56.01	71.67	100.00

Table B1a: Contribution of Changing Moral Attitudes and Negotiating IEAs to Reducing Global Emissions

Table B1b shows the outcomes in terms of welfare. The effects on welfare are smaller than for emissions. In the cases where $\kappa_\varepsilon = \kappa_\mu$, for the lowest value 0.1, forming an IEA adds slightly less than changing individual moral attitudes; but for the highest value, 0.5, changing individual moral attitudes contributes three-quarters of the overall benefit of forming the grand coalition.

κ_ε	Imperfect Kantian Non-Coop Equilibrium	IEAs- κ_μ		
		0.1	0.3	0.5
0.1	19.00	35.13	66.15	100.00
0.3	51.00	60.76	79.53	100.00
0.5	75.00	79.98	89.55	100.00

Table B1b: Contribution of Changing Moral Attitudes and Negotiating IEAs to Raising Global Welfare

This competes our analysis of the model with a linear damage cost function.

Appendix C: Equilibrium Outputs of Model with Quadratic Benefit Function and Linear Damage Cost Function.

From the definition of *B*(.) and *D*(.) it is straightforward to derive the following.

First-best (Fully Cooperative) and Self-Interested Non-Cooperative Outcomes

$e^{SO} = e^{FC} = \beta - \delta n;\ \ W^{SO} = W^{FC} = 0.5(e^{*})^2;$

$e^{NC} = e^{SO} + \delta(n-1);\ \ W^{NC} = W^{SO} - 0.5[\delta(n-1)]^2$

Kantian Non-Cooperative Outcomes

$\hat{e}(\kappa) = e^{SO} + \delta(1-\kappa)(n-1);\ \ \widehat{W}(\kappa) = W^{SO} - 0.5[\delta(1-\kappa)(n-1)]^2$

$\hat{\hat{e}}(\kappa) = \kappa;\ \ \hat{\hat{e}}'(.) > 0;\ \ \widehat{\widehat{W}}(\kappa) = 1-(1-\kappa)^2;\ \ \widehat{\widehat{W}}'(.) > 0\ ;$

$\hat{\hat{e}}(0) = \widehat{\widehat{W}}(0) = 0;\ \hat{\hat{e}}(1) = \widehat{\widehat{W}}(1) = 1$

IEAs with Kantian Behaviour – Stage 2 Equilibrium Emissions, Payoffs and Welfare

$\tilde{e}^{c}(m,\kappa_\varepsilon) = e^{SO} + \delta(1-\kappa_\varepsilon)(n-m);\ \tilde{e}^{f}(m,\kappa_\varepsilon) = e^{SO} + \delta(1-\kappa_\varepsilon)(n-1)$

$\tilde{e}^{a}(m,\kappa) = e^{SO} + \left(\frac{\delta}{n}\right)[(1-\kappa)(n-m)(n+m-1)];$

$\tilde{\Pi}^{c}(m,\kappa_\varepsilon,\kappa_\mu) = W^{SO} - 0.5\delta^2(1-\kappa_\varepsilon)^2(1-\kappa_\mu)[(n-1)^2-(m-1)^2]$

$\tilde{\Pi}^{f}(m,\kappa_\varepsilon,\kappa_\mu) = W^{SO} - 0.5\delta^2(1-\kappa_\varepsilon)^2[(n-1)^2-2(1-\kappa_\mu)m(m-1)]$

$\widetilde{W}^{c}(m,\kappa) = W^{so} - 0.5\delta^2(1-\kappa)[(1-\kappa)(n-m)^2 + 2(n-m)(m-1)];$

$\widetilde{W}^{f}(m,\kappa) = W^{so} - 0.5\delta^2(1-\kappa)[(1-\kappa)(n-1)^2 - 2m(m-1)];$

$\widetilde{W}^{a}(m,\kappa) = W^{so} - 0.5\delta^2(1-\kappa)^2(n-m)\left[\frac{m(n-m)+(n-1)^2}{n}\right];$

IEAs with Kantian Behaviour – Stage 1

$\sigma(m,\kappa_\varepsilon,\kappa_\mu) = 0.5\delta^2(1-\kappa_\varepsilon)^2[\kappa_\mu(n-1)^2 - (1-\kappa_\mu)(m-1)(m-3)]$

i.e. $\sigma(m,\kappa_\varepsilon,\kappa_\mu) = 0.5\delta^2(1-\kappa_\varepsilon)^2(1-\kappa_\mu)[\chi-(m-1)(m-3)]$, where $\chi = \frac{\kappa_\mu(n-1)^2}{(1-\kappa_\mu)}$.

Clearly the sign of $\sigma(.)$, and hence the size of the stable IEA, does not depend on either δ or $\kappa_\varepsilon, 0 \le \kappa_\varepsilon < 1$. $2+\sqrt{1+\chi}$ is the real value of *m* for which $\sigma(m,\kappa_\varepsilon,\kappa_\mu) = 0$. Therefore, the size of the stable IEA is $\widetilde{m}(\kappa_\mu)$ - the largest integer no greater than $\min(n, 2+\sqrt{1+\chi})$. It is straightforward to see that $\widetilde{m}(\kappa_\mu)$ is increasing in κ_μ and that $\widetilde{m}(0.5,n) = n$.

Equivalently, define $\underline{\kappa}_\mu(m)$ as the minimum value of κ_μ for which *m* is a stable IEA, i.e. $\sigma(m,\kappa_\varepsilon\underline{\kappa}_\mu) = 0$; so $\underline{\kappa}_\mu(m) = \frac{(m-1)(m-3)}{(n-1)^2+(m-1)(m-3)}$, $\underline{\kappa}_\mu(n) = \frac{(n-3)}{2(n-2)} < 0.5$.

IEAs with Kantian Behaviour – Overall Equilibrium

$$\tilde{\tilde{e}}^c(\kappa_\varepsilon, \kappa_\mu) = 1 - \frac{(1-\kappa_\varepsilon)\left(n-\tilde{m}(\kappa_\mu)\right)}{(n-1)}; \quad \frac{\partial \tilde{\tilde{e}}^c}{\partial \kappa_\varepsilon} > 0; \quad \frac{\partial \tilde{\tilde{e}}^c}{\partial \kappa_\mu} > 0 \; ; \quad \tilde{\tilde{e}}^f(\kappa_\varepsilon) = \kappa_\varepsilon; \quad \frac{\partial \tilde{\tilde{e}}^f}{\partial \kappa_\varepsilon} > 0$$

$$\widetilde{W}^c(\kappa_\varepsilon, \kappa_\mu) = W^* - 0.5\delta^2(1-\kappa_\varepsilon)(n - \tilde{m}(\kappa_\mu)[(1-\kappa_\varepsilon)\left(n - \tilde{m}(\kappa_\mu)\right) + 2\left(\tilde{m}(\kappa_\mu) - 1\right)]$$

$$\widetilde{W}^f(\kappa_\varepsilon, \kappa_\mu) = W^* - 0.5\delta^2(1-\kappa_\varepsilon)[(1-\kappa_\varepsilon)(n-1)^2 - 2\tilde{m}(\kappa_\mu)\left(\tilde{m}(\kappa_\mu) - 1\right)]$$

$$\tilde{\widetilde{W}}^c(\kappa_\varepsilon, \kappa_\mu) = 1 - \frac{(1-\kappa_\varepsilon)(n-\tilde{m}(\kappa_\mu))[(1-\kappa_\varepsilon)\left(n-\tilde{m}(\kappa_\mu)\right)+2\left(\tilde{m}(\kappa_\mu)-1\right)]}{(n-1)^2}$$

$$\tilde{\widetilde{W}}^f(\kappa_\varepsilon, \kappa_\mu) = 1 - \frac{(1-\kappa_\varepsilon)[(1-\kappa_\varepsilon)(n-1)^2 - 2\tilde{m}(\kappa_\mu)\left(\tilde{m}(\kappa_\mu) - 1\right)]}{(n-1)^2}$$

$$\tilde{m}(\kappa_\mu) = 1 \Longrightarrow \tilde{\widetilde{W}}^f(.) = 1 - (1-\kappa_\varepsilon)^2; \; \tilde{m}(\kappa_\mu) = n \Longrightarrow \tilde{\widetilde{W}}^c(.) = 1;$$

$$\frac{\partial \tilde{\widetilde{W}}^c}{\partial \kappa_\varepsilon} > 0; \; \frac{\partial \tilde{\widetilde{W}}^f}{\partial \kappa_\varepsilon} > 0; \; \frac{\partial \tilde{\widetilde{W}}^f}{\partial \kappa_\mu} > 0; \; \frac{\partial \tilde{\widetilde{W}}^c}{\partial \kappa_\mu} = \frac{2(1-\kappa_\varepsilon)(1+\kappa_\varepsilon)\tilde{m}'}{(n-1)^2}[\tilde{m}(\kappa_\mu) - \frac{1+n\kappa_\epsilon}{1+\kappa_\varepsilon}]$$