Economic Studies
at BROOKINGS

ES Working Paper Series, September 2014

Optimal Altruism in Public Good Provision

Robert W. Hahn, Smith School of Enterprise and the Environment, Oxford University Institute for New Economic Thinking (INET)
Robert A. Ritz, Faculty of Economics, Cambridge University Energy Policy Research Group (EPRG)

Abstract

We present a model of altruistically-minded—yet rational—players contributing to a public good. A key feature is the tension between altruism and crowding-out effects. We present three main results: (1) More altruistic behaviour often reduces social welfare; (2) It is almost always optimal for a player to act more selfishly than her true preference; (3) A player's optimal altruistic commitment is often low or zero—even with strongly altruistic preferences. Applications to a range of public good problems, including climate policy, are discussed. Our results highlight that it will generally be difficult to infer social preferences from observed behaviour.

We would like to thank Toke Aidt, David Anthoff, Elizabeth Baldwin, John Feddersen, Reyer Gerlagh, Thomas Greve, Cameron Hepburn, Charles Mason, Grischa Perino, Rick van der Ploeg, John Quah, Robert Stavins, Paul Tetlock, Alexander Teytelboym, Richard Tol, and Alistair Ulph for helpful comments and advice, seminar participants at EPRG, OxCarre, Cambridge, and Toulouse for discussions. The usual disclaimer applies.

1 Introduction

There is a growing recognition that social preferences may play an important role in explaining economic outcomes such as those arising in problems of public good provision.[1] We study the welfare impact of unselfish behaviour by altruistically-minded—yet rational—players, and ask to what extent a preference for altruism is optimally reflected in a player's contribution to a public good. To our knowledge, this is the first attempt in the literature to understand a notion of "optimal altruism".

Our analysis is motivated in part by recent experience with climate policy, which many consider to be one of the biggest public good problems of today (Stern 2008). Recent years have witnessed a number of unilateral initiatives to combat climate change at the local, national, and regional levels. For example, the EU has a program to reduce greenhouse gas emissions by 20% (relative to 1990 levels) by 2020 while the UK aims to cut emissions by 80% by 2050.[2] Such initiatives have taken place in the absence of a global agreement by countries to jointly reduce emissions, e.g., with a global cap-and-trade scheme.

Relatedly, there is an increasing use of the "social cost of carbon" (SCC) in regulatory decision-making. The SCC reflects the marginal benefit to the world from reducing CO_2 emissions—rather than only to an individual country or region. Several European countries have applied the SCC (Watkiss and Hope 2012), and the US has also developed a measure of the SCC (Greenstone, Kopits and Wolverton 2013) which to date has been applied to selected energy and environmental regulations. At the same time, many other countries do not incorporate the SCC in policymaking, and do not appear to have engaged in emissions abatement beyond "business-as-usual".

There is some evidence that the domestic costs associated with unilateral policies exceed domestic benefits. For example, Tol's (2012) cost-benefit analysis of the European Union's 20/20/20 policy package finds a benefit-cost ratio < 1 across a range of scenarios.[3,4] In a similar vein, the UK Department of Energy and Climate Change's impact assessment of the 2008 Climate Change Act finds "the economic case for the UK continuing to act alone where global action cannot be achieved would be weak" (DECC 2009).

It is difficult to reconcile these unilateral initiatives with standard economic theory, including the theory of international environmental agreements (Barrett 1994, 2005). Put simply, if unilateral action by local, national, or regional actors reduces their own domestic welfare, then why are they doing it? But it seems possible that some of these climate initiatives may be a reflection of "unselfish" or "altruistic" motives, in the sense of incor-

[1] See Sobel (2005) for an overview of interdependent preferences in economic analysis.

[2] Similiar climate-policy initiatives, many on a relatively small scale, also exist, for example, in Australia, California, China, Japan, New Zealand and Norway, as well as at the city level.

[3] This policy package targets a 20% cut in greenhouse gas emissions, a 20% share of renewable energy, and a 20% improvement in energy efficiency by 2020.

[4] Similar issues have also emerged in the analysis of the recently proposed regulation to reduce carbon dioxide emissions from the US power sector (EPA 2014). The economic analysis conducted by the EPA suggests that the global benefits of this regulation exceed its costs but, in some scenarios, domestic benefits accruing only to the US fall short of domestic costs (e.g., depending on the extent to which health-related benefits are taken into account in addition to direct climate benefits).

porating benefits that accrue outside the borders of the acting jurisdiction. This paper seeks to understand the role that altruism can play in such public goods problems.[5]

We begin with a two-player model of (non-cooperative) public good provision with the following key features. A player's net benefit or "national welfare" Π_k $(k = i, j)$ equals the benefit she derives from total contributions to the public good (by both players) minus the cost of her own contribution, while "global welfare" $W = \Pi_i + \Pi_j$.[6] Preferences may depart from self-interest: A player's true objective function $S_k = (1-\theta_k)\Pi_k + \theta_k W$ places weight on both her own net benefits and global welfare, where $\theta_k \in [0, 1]$ represents her degree of altruism.[7] More altruistic behaviour by player i leads to an increase in its own public good contribution but induces player j to cut back ("crowding out"). We refer to the rate at which the other player's effort contracts as the "leakage" rate. The tension between altruism and leakage lies at the core of our analysis. Our modelling approach is also consistent with a characteristic shared by many (global) public good problems: The absence of a world government means that solutions enforced by a central mechanism designer play a limited role.[8]

Our analysis highlights three main findings. First, we obtain the seemingly paradoxical result that *more altruistic behavior by an individual player often reduces social welfare.* For example, consider a small commitment to more altruistic behavior by player i. Such a commitment raises the equilibrium net benefit enjoyed by player j but reduces i's own net benefit. We show that welfare is more likely to fall if player i derives an above-average marginal benefit from contributions, and the leakage rate from her commitment is higher.[9] Conversely, a necessary condition for more altruistic behaviour to raise such a player's true objective is that her degree of altruism exceeds her leakage rate. This already shows that whether altruism is privately optimal and/or welfare-augmenting depends crucially on

[5] Our analysis focuses on "international altruism" between countries rather than "intergenerational altruism" between different generations of people in a single economy. We differ from much of the literature on altruism in that we often think of our unit of analysis as a country rather than an individual. Also, we do not wish to claim that social preferences are the only possible way of explaining unilateral climate action; in some cases, other explanations, e.g., domestic political economy, may be important.

[6] Our results are robust to different welfare definitions. Section 5 provides details.

[7] Our formulation of altruism has a continuum of preferences, ranging from entirely selfish to entirely altruistic preferences. On a historical note, Edgeworth (1881) uses essentially the same formulation, by writing $S_i = \Pi_i + \theta_i \Pi_j$ and calling θ_i the "coefficient of effective sympathy". Some other formulations of altruism have conditional elements. For example, the models of inequity aversion due to Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) feature utility functions with reference points which determine the degree of perceived inequity in payoffs (and also affect players' actions, e.g., depending on whether they are "ahead" or "behind"). Lange and Vogt (2003) show that a preference for equity can generate cooperation in international environmental negotiations, while Kosfeld, Okada and Riedl (2009) argue that fairness can play an important role in the formation of institutions geared towards improving public good provision. See also Rabin (1993) on fairness in economic analysis.

[8] Key contributions on voluntary public-good provision include Bergstrom, Blume and Varian (1986) and Cornes and Sandler (1996). We work with a simplified model with reduced-form benefit and cost functions, as is standard in much of the environmental economics literature (e.g., Hoel 1991; Barrett 1994), which captures the key feature that players' contributions are strategic substitutes.

[9] Hoel (1991) obtains a related result in an important early model of unilateral commitment in environmental policy that does not feature social preferences. He shows that a small (exogenous) commitment to a higher public good contribution by a player, starting from a world in which all players make entirely selfish contributions, can reduce global welfare. Our work goes further by examining a world in which players can behave altruistically to different degrees, and deriving a notion of optimal (endogenous) altruism.

the details of the environment; a player may thus wish to find ways of making public contributions that departs from her true objective.

Second, we show that *a player who genuinely wants to maximize global welfare almost always does best by being at least somewhat selfish.* To see this, suppose that player i's true preference is entirely altruistic, $\theta_i = 1$, while player j is altruistic only to some degree, $\theta_j < 1$. Should i make the contribution that maximizes its underlying global-welfare objective? No. Intuitively, a small decrease in its own contribution only leads to a second-order loss in global welfare (by the envelope theorem). But the resulting induced *increase* in the other player's effort leads to a first-order gain (whenever the other player is not already choosing the first-best effort).[10] This is what we call "reverse leakage"—a weaker commitment reduces free-riding by other players, and this can raise social welfare.

Third, we find that *a player's optimal altruistic commitment is often "low" or zero—even with strongly altruistic preferences.* In some cases, it is optimal for a player who cares about global welfare to act entirely selfishly, maximising *only* her own net benefit. We thus highlight that caution is required in inferring whether or not players are "being selfish" from their observed behaviour; selfish behaviour may be a welfare-maximising response to crowding-out effects, especially with heterogeneous players.

We characterize optimal altruistic commitments using the following modelling device: Player k has a strategic objective function $\Omega_k = (1 - \lambda_k)\Pi_k + \lambda_k S_k$, where $\lambda_k \geq 0$ is her strategic preference. A player chooses a public good contribution according to her true preference if $\lambda_k = 1$, but whenever $\lambda_k < 1$ ($\lambda_k > 1$) acts more (less) selfishly than would be her true preference. We determine a player's optimal commitment $\lambda_k^*(\theta_i, \theta_j)$ to incorporate its altruism into public good contribution. In particular, we always have $\lambda_k^* \leq 1$, almost always find $\lambda_k^* < 1$ (for $k = i, j$), and, in a range of cases, $\lambda_i^* \approx 0$ and/or $\lambda_j^* \approx 0$.[11] Only where *all* players have entirely altruistic preferences $\theta_i = \theta_j = 1$, is a full commitment $\lambda_k^* = 1$ (for $k = i, j$) optimal, in which case the first-best outcome obtains.

We show that these results are very robust to a variety of different model specifications. This includes the generalization to $n \geq 3$ players—where we exploit the fact that players' contributions are made in an "aggregative game" (in the sense of Corchón 1994; see also Cornes and Sandler 2007); moderate degrees of cross-country cost spillovers (e.g., in renewable energy technologies such as solar or wind); and different representations of altruism in players' objective functions, including the "warm glow" of Andreoni (1989, 1990). Thus, our results apply with both "pure" and "impure" forms of altruistic preferences.

One way of thinking about how a player can commit to actions that depart from her true preference is in terms of the theory of strategic delegation. For example, citizens may delegate decision-making on abatement targets to politicians, and may wish to appoint politicians whose climate-policy preference differs from their own (e.g., from those of the

[10] These basic insights rely on crowding-out effects but not on whether leakage rates are "high" or "low".

[11] To illustrate, two countries' true preferences may be to apply the global SCC to 100% and 46% of projects respectively, that is, $(\theta_i, \theta_j) = (1, \frac{6}{13})$. But if $(\lambda_i^*, \lambda_j^*) = (\frac{1}{2}, 0)$, say, then optimal altruism involves using the SCC only in 50% ($= \frac{1}{2} \times 1$) of projects for country i and not at all for country j ($= 0 \times \frac{6}{13}$). (The details underlying this numerical example are at the end of Section 6.)

median voter). Commitment can also be achieved by political or regulatory institutions—perhaps independent of government—which adopt particular rules and practices.

The classic reference on such delegation is Schelling (1960), and the idea has been applied widely to different contexts such as bargaining (Segendorff 1998), monetary policy (Persson and Tabellini 1993), and the theory of the firm (Vickers 1985). It is fairly well-known that an incentive to misrepresent preferences exists in virtually any game (Heifetz, Shannon and Spiegel 2007)—although this, in itself, says little about *how* preferences will be distorted in a particular game.

We differ from this literature in several respects. To begin with, we consider a different class of game, and examine a setting in which agents are not driven by pure self-interest; many of our themes thus have no analog in previous models.[12] Moreover, our main application to climate policy has at least two advantages compared to other delegation applications. First, there is significant empirical evidence that players' efforts are strategic substitutes: A very large majority of work on unilateral climate policy finds that carbon leakage rates are positive, as in our model.[13] Second, climate policy is characterized by something close to an "informational level playing field" between countries: The climate-change debate is highly public and global (based, in part, on scientific evidence) and countries' abatement policies are commonly known (perhaps with a few exceptions), as is whether or not they have adopted the SCC.[14]

Our analysis also shows that altruism can, at least in principle, *neutralize* the strategic incentive to distort preferences which is emphasized by this literature. Our model always features strategic substitutes, so contributing less induces a favourable response from the other player; but if both players are fully altruistic, they recognize that such preference distortion no longer yields any gain (as it induces a move away from first-best). So an incentive to distort play may exist in a standard game with selfish players—but not in an otherwise identical game featuring social preferences.

Other applications. Questions of altruistic behaviour arise in other environmental problems. For example, there is an ongoing debate about the motivations behind the Montreal Protocol to reduce chlorofluorocarbons (CFCs) which deplete the ozone layer. While Barrett (1994) argues that the protocol was broadly consistent with the outcome of a

[12] Perhaps closest to us, though in a rather different setup without altruism, Roelfsema (2007) considers a model of imperfect competition with strategic trade policies, in which delegation to a politician who cares more about the environment than the median voter can be optimal because this induces other countries to do the same. (This result relies on a particular form of competition in product markets.)

[13] Many empirical estimates are derived from numerical simulations of multi-sector, general equilibrium models which focus on climate initiatives by OECD countries that result in carbon leakage to non-OECD countries. These typically find leakage in the range of 5–40%, with many estimates below 20%. Leakage estimates for individual sectors (such as the cement and steel industries in the EU Emissions Trading Scheme) are frequently higher, e.g., above 50%, but also rarely exceed 100%. See Babiker (2005), Copeland and Taylor (2005), Ritz (2009), and the references cited therein.

By contrast, in delegation models on the theory of the firm, it is often difficult to tell with confidence whether competition between firms is in strategic substitutes (Cournot) or in strategic complements (Bertrand)—and many results are known to depend critically on this unobservable feature of the model.

[14] In practice, there is significant uncertainty over the costs and benefits of CO_2 abatement; the key point for us is that actions are easily observable and informational asymmetries between countries are small.

non-cooperative game, Sunstein (2007) notes that the US used a relatively low discount rate in evaluating its commitment—which *might* be interpreted as a form of altruism.

The key features of our model are shared by other problems of the commons. In fisheries policy, for example, there is a strong tendency towards overexploitation; individual players have a suboptimal incentive to limit their catch (Stavins 2011) and catch reductions are typically strategic substitutes (Levrahi and Mirman 1980), leading to a leakage problem analogous to ours. Similarly, in a classic paper, Olson and Zeckhauser (1966) suggest that small countries tend to free-ride on the defense investments of large countries, and observe that countries' military expenditures are often strategic substitutes. It is more difficult to pinpoint altruism empirically in these applications, partly because there is no clear equivalent to the adoption of the SCC. However, it seems conceivable that individual European countries, say, also care about the welfare of the EU as a whole when it comes to policies affecting the environment or defense.[15]

Our results can also apply to problems from other domains that share public good characteristics. For example, suppose family member j pursues some useful activity; family member i derives indirect benefits from the activity, and can help out at some cost. If altruistic, i also cares about the benefits accruing to j in choosing how much to help. But the more i helps, the less j does himself—the leakage problem. While i's help always raises j's private payoff, it need not raise overall welfare or i's own altruistic objective. Optimal altruism typically involves $\lambda_i^* < 1$, so i's help falls short of her true preference.[16] To be concrete, a parent may want to help a child with its homework on 4 out of 5 days a week ($\theta_i = \frac{4}{5}$, say), but realizes that, because of incentive effects, $\lambda_i^* = \frac{1}{2}$, say, is optimal—and thus only helps twice a week. In practice, such a well-meaning but stern commitment may be achieved by putting certain rules into place, or the parent may engage a tutor (or sibling) twice a week and abstain from helping directly.[17]

Plan for the paper. Section 2 sets up our benchmark model. Section 3 examines the impact of "small" altruistic commitments. Section 4 analyzes in detail players' optimal commitments, and Section 5 shows that our main results are robust in a variety of directions. Section 6 points out some further properties of our model, with a focus on its empirical implications. Finally, Section 7 discusses recent climate policy initiatives in light of our results, and offers some suggestions for future research. (The proofs are in Appendix A, and the details of the robustness analysis are in Appendix B.)

[15] A related application is the problem faced by large charities like the Bill & Melinda Gates or Rockefeller foundations. It seems clear that the broad objective of such organizations is to enhance some measure of global welfare. At the same time, there are well-known concerns that their contributions can "crowd out" others, such as local governments, the private sector, and smaller charities. This corresponds quite directly to the tension between altruism and leakage in our analysis.

[16] The "rotten kid theorem" (Becker 1974) does not apply in our model. It states that, under certain conditions (Bergstrom 1989), an altruistic head who makes transfers to self-interested household members induces the efficient outcome—despite limited altruism in the family overall. By contrast, our setup does not feature a design with transfer payments (see also our concluding discussion in Section 7). (Recall also that the rotten kid theorem itself can fail in public-good settings, despite transfers.)

[17] Our model assumes the tutor is optimally chosen and incentivized by the parent (strictly speaking, at zero cost) and has no special skills—although this is clearly not essential for the results.

2 A model of altruism in public good provision

Setup of the model. Two players, i and j, contribute to the provision of a public good. Player k ($k = i, j$) makes a contribution (e.g., shared investment, emissions reduction, or "effort") denoted by X_k, and derives benefits $B_k(X_i + X_j)$ which depend on the aggregate effort by the two players. The marginal benefit satisfies $B_k'(\cdot) > 0$ and $B_k''(\cdot) < 0$. The cost function $C_k(X_k)$ is player-specific, with marginal cost satisfying $C_k'(\cdot) > 0$ and $C_k''(\cdot) > 0$. To guarantee an interior solution, assume $C_k(0) = C_k'(0) = 0$ and $B_k'(\overline{X}_k) - C_k'(\overline{X}_k) < 0$ for some $\overline{X}_k < \infty$. Define a player k's "net benefit" or "national welfare" as $\Pi_k = B_k(X_i + X_j) - C_k(X_k)$, and "social surplus" or "global welfare" as $W = \Pi_i + \Pi_j$.

In our model, each player's preferences may be at least partly altruistic. In particular, player k's *true* objective function is given by

$$S_k = (1 - \theta_k)\Pi_k + \theta_k W, \tag{1}$$

where the parameter $\theta_k \in [0, 1]$ represents her true preference for altruism. Player k is purely self-interested if $\theta_k = 0$ (so $S_k = \Pi_k$), and entirely altruistic if $\theta_k = 1$ (so $S_k = W$), in which case her preference reflects the full *global* benefit of contributions, $B_i + B_j$. More generally, a higher value of θ_k represents a "more altruistic" preference that gives more weight to the other player's net benefit. For our application to climate policy, we can interpret $\theta_k = 0$ as an underlying preference for the "business-as-usual" (BAU) level of emissions, while $\theta_k = 1$ corresponds to a desire to incorporate the global "social cost of carbon" (SCC) into decision-making.

We next introduce a modelling device in form of a *strategic* objective function:

$$\Omega_k = (1 - \lambda_k)\Pi_k + \lambda_k S_k. \tag{2}$$

A strategic objective is a convex combination of a player's net benefit Π_k and her true objective S_k, with a relative weight given by the strategic preference $\lambda_k \in [0, \theta_k^{-1}]$. If $\lambda_k = 0$, the strategic objective is entirely selfish, so $\Omega_k = \Pi_k$ (regardless of the underlying true objective S_k). If $\lambda_k = 1$, the player's strategic objective is identical to her true objective, so $\Omega_k = S_k$. We restrict attention to $\lambda_k \leq \theta_k^{-1} \Leftrightarrow \lambda_k\theta_k \leq 1$ to focus on the typical situation where each player contributes too little to the public good from a social-welfare perspective—rather than too much. (Whenever $\theta_k < 1$, we do allow for the possibility that $\lambda_k > 1$ so the strategic objective *could* place more weight on altruism than the true objective—although we will see that, in equilibrium, this does not occur.) This (λ_i, λ_j)-modelling device allows us to analyze the welfare impact of players following through on their altruistic preferences, and, building on this, to understand the extent to which players optimally engage in altruistic behaviour.[18]

The timing of the model is as follows. At Date 0, each player is endowed with a

[18] In the literature on international environmental agreements, countries typically make a binary decision on joining an agreement ("in or out"). By contrast, countries here choose the intensity of their commitment.

benefit function and a cost function, $B_k(\cdot)$ and $C_k(\cdot)$, as well as with a true objective $S_k(\cdot)$ that reflects her degree of altruism, $\theta_k \in [0,1]$. Then, at Date 1, each player chooses her strategic preference $\lambda_k \in [0, \theta_k^{-1}]$ to maximize her true objective S_k. Finally, at Date 2, each player—or her agent—chooses effort according to the strategic objective function Ω_k. (For environmental applications, a country's choice of X_k is equivalent to choosing a domestic price on emissions.[19])

We focus on the subgame-perfect Nash equilibrium of the game, and follow the delegation literature in assuming that players' strategic objective functions, Ω_i and Ω_j, form credible commitments.[20] The plausibility of this assumption will, of course, vary depending on the application in question. As explained in the introduction, we think that commitment value is reasonably likely to obtain in the climate-policy context, given something close to an informational level playing field between countries, as well as in other public good problems.

Key properties of the model. We begin by establishing the key properties of the model at Date 2. For player i, say, the first-order condition for its contribution is

$$\partial\Omega_i/\partial X_i = (B_i' - C_i') + \lambda_i\theta_i B_j' = 0. \tag{3}$$

The "first-best" benchmark is nested where (i) both players have entirely unselfish true preferences, $\theta_i = \theta_j = 1$, *and* (ii) both players choose their respective effort levels accordingly, $\lambda_i = \lambda_j = 1$. In this case, players at Date 2 make contribution decisions to $\max_{X_k} W$ (where $k = i, j$), thus each incorporating the full global benefit of their actions.

The first-order condition also defines player i's best response to player j's contribution, $R_i(X_j)$. The slope of this function is given by

$$R_i'(X_j) = \frac{\left(B_i'' + \lambda_i\theta_i B_j''\right)}{(-B_i'' + C_i'' - \lambda_i\theta_i B_j'')} \in (-1, 0). \tag{4}$$

A key property of the model is that *players' efforts are strategic substitutes.* This captures a "crowding out" effect: If one player increases her effort, this reduces the marginal benefit of effort for the other player, who therefore responds by cutting back. In the context of climate policy, $L_i \equiv [-R_j'(X_i)] \in (0,1)$ is the marginal rate of "*carbon leakage*" (IPCC, 2007) resulting from country i's effort. Borrowing this terminology:

Lemma 1 *The leakage rate due to player k's effort is given by $L_k \in (0,1)$.*

[19] To see this, imagine splitting country k's abatement decision at Date 2 into two parts. At Date 2b, a representative, price-taking firm chooses emissions abatement X_k to maximize its profits $p_k X_k - C_k(X_k)$, where p_k is the domestic emissions price, such that $p_k = C_k'(X_k)$, in equilibrium. This defines an upward-sloping abatement supply curve with $dX_k/dp_k = 1/C_k''(X_k) > 0$. At Date 2a, policymakers choose the domestic price p_k to maximize the strategic objective Ω_k. This setup is exactly equivalent to the benchmark model since choosing the domestic emissions price is equivalent to choosing an abatement effort.

[20] This is essentially equivalent to assuming that players' contributions are publicly observable, which, in turn, corresponds to i knowing j's $\lambda_j\theta_j$ when choosing her contribution policy at Date 2 (as benefit and cost functions are commonly known). (We do not require that i then knows j's true preference θ_j.)

Leakage rates quantify the severity of the crowding-out problem; they are positive but less than 100%. This is a common feature of public good models across different domains, including environmental problems, military protection, fisheries, and charitable giving.

We next confirm the intuition that more altruistic behaviour by a player leads to an increase in her effort. (The result from Lemma 1 ensures that the equilibrium is unique, stable, and exhibits well-behaved comparative statics.)

Lemma 2 *If player k's true preference $\theta_k > 0$, her effort satisfies $dX_k^*/d\lambda_k > 0$.*

A higher value of λ_k inflates the marginal return to public good contribution, which, by stability, also increases its equilibrium level. So an increase λ_i, say, raises X_i^* (Lemma 2) and also raises $X_i^* + X_j^*$, but not by as much (Lemma 1).

To complete our preliminary discussion, we show that a player with an entirely selfish true preference, $\theta_k = 0$, does not want to engage in a strategic commitment.

Lemma 3 *If player k's true preference $\theta_k = 0$, her optimal effort solves $\max_{X_k} \Pi_k$.*

As a notational convention, we refer to such an optimal commitment as $\lambda_k^* = 0$.[21]

3 The welfare impact of small altruistic commitments

To build intuition, we begin our analysis by considering "small" commitments. Suppose that player i's true preference $\theta_i > 0$ is altruistic at least to some extent (while $\theta_j \geq 0$), and that initially both players act purely in their self-interest, i.e., $\lambda_k = 0$ for $k = i, j$. What is the impact of a small commitment $d\lambda_i > 0$ by player i towards incorporating her true altruistic preference in her public good contribution?

Proposition 1 *The impact of a small unilateral commitment $d\lambda_i > 0$ by player i on her equilibrium true objective*

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_i=\lambda_j=0} = \left[\left(\theta_i B_j' - B_i' L_i\right)\frac{dX_i^*}{d\lambda_i}\right]_{\lambda_i=\lambda_j=0}$$

is (a) positive if the ratio of marginal benefits satisfies $B_i' \leq B_j'$ and her true preference exceeds the leakage rate $\theta_i > L_i$, and (b) negative for a ratio of marginal benefits B_i'/B_j' sufficiently large or for a true preference θ_i sufficiently small.

Whether a small altruistic commitment is beneficial for a player depends crucially on the details of the environment. If she either derives a relatively large marginal benefit, or her true preference contains only a small degree of altruism, it is never a good idea for someone to make such a commitment. However, two simple conditions which are jointly

[21] This convention makes an altruistic player's strategic commitment directly comparable with a selfish player; note that, if $\theta_k > 0$, then the contribution solves $\max_{X_k} \Pi_k$ if and only if $\lambda_k = 0$. (Recall that we are restricting attention to cases where $\lambda_k \geq 0$.)

sufficient for $dS_i^* > 0$ are that the player has a relatively low marginal benefit as well as a true preference that exceeds the rate of leakage.

These results can be understood as follows. With a slight abuse of notation, let $dX_i^* > 0$ denote the increase in i's effort due to its small unilateral commitment $d\lambda_i > 0$. (More formally, $dX_i^* = \left[\left(dX_i^*/d\lambda_i\right)_{\lambda_i=\lambda_j=0}\right] d\lambda_i > 0$ by Lemma 2.) Due to the crowding-out effect (Lemma 1), j adjusts its effort by $dX_j^* = (-L_i)\, dX_i^* < 0$ in response. By the envelope theorem, the *direct* effect of a small change in each player's effort on its *own* net benefit is zero. The reason is that both players were initially choosing their respective efforts selfishly to maximize their own net benefit, so any (small) change their own contribution only has a second-order effect. However, the unilateral commitment by i also has two *strategic* effects, one positive and one negative. First, the increase in i's effort yields an increase in the benefits enjoyed by the *other* player j of $B_j' dX_i^* > 0$. Second, the induced reduction in j's effort means that i's benefit changes by $B_i' dX_j^* = (-B_i' L_i)\, dX_i^* < 0$. Player i's true objective $S_i = \Pi_i + \theta_i \Pi_j$ places weight $\theta_i \in [0, 1]$ on the first (positive) strategic effect and full weight on the second (negative) strategic effect. The weighted sum of these effects, $(\theta_i B_j' - B_i' L_i) dX_i^*$, thus determines the impact of a small unilateral commitment on its own true objective function and behaves according to Proposition 1.

Intuitively, a unilateral commitment by i increases the net benefit Π_j^* enjoyed by j but acting unselfishly hurts its own net benefit Π_i^*. The commitment thus enhances its own true objective if (and only if) the former effect outweighs the latter. The positive effect will be large if j's marginal benefit is large, and receives large weight according to i's degree of altruism, θ_i. The negative effect will be small if there is little leakage, and if i's own marginal benefit is small.

We can also address when a small altruistic commitment improves *global* welfare:

Proposition 2 *The impact of a small unilateral commitment $d\lambda_i > 0$ by player i on equilibrium global welfare*

$$\left.\frac{dW^*}{d\lambda_i}\right|_{\lambda_i=\lambda_j=0} = \left[\left(B_j' - B_i' L_i\right) \frac{dX_i^*}{d\lambda_i}\right]_{\lambda_i=\lambda_j=0}$$

is (a) positive if the ratio of marginal benefits satisfies $B_i' \leq B_j'$, and (b) negative for a ratio of marginal benefits B_i'/B_j' sufficiently large.

The logic underlying Proposition 2 follows that of Proposition 1. Again, the direct effects on each player's net benefit are both zero by the envelope theorem. The only difference arises because, from a global-welfare perspective, the combined effect of the two strategic effects depends on their *unweighted* sum. So the increase in the benefits enjoyed by the *other* player j of $(B_j')dX_i^* > 0$ plus the induced reduction in i's benefit of $[B_i'(-L_i)]dX_i^* < 0$ yield an overall welfare impact $dW^* = (B_j' - B_i' L_i)dX_i^*$. The sign of this expression, too, is ambiguous. However, note that Proposition 2(a) implies that a small commitment must be global welfare-enhancing for at least one of the two players.

A small commitment, if it occurs, is of course more likely to raise global welfare than i's true objective. For example, with identical benefit *functions*, $B_i(X_i + X_j) = B_j(X_i + X_j)$, i's small commitment always raises equilibrium global welfare W^* (Proposition 2, since $L_i < 1$) but its own true objective S_i^* may still decline (Proposition 1, for $\theta_i < L_i$).[22]

4 Optimal altruistic commitments

This section develops our main results on "optimal altruism". Our analysis so far has already shown that a more altruistic commitment can raise or reduce social welfare; this already suggests that players may wish to make public good contributions in ways that depart from their true objectives.

We begin by deriving a generalized formula for the welfare impact of more altruistic behaviour, and then use this to establish our main arguments. First, we show that optimal altruism *almost* always means that players act more selfishly than would be their true preference. Second, optimal commitments are often *much* lower than players' true preferences and, in a range of cases, a socially-concerned player does *best* by acting entirely selfishly. Thus it will generally be difficult to empirically infer social preferences from observed behaviour.

The general model. In the general version of the model, each player chooses optimally how altruistically to act so as to maximize her true objective S_k; this yields equilibrium values $\lambda_k^*(\theta_i, \theta_j)$ for players' strategic preferences ($k = i, j$). This analysis is more complicated because our previous argument, based on the envelope theorem, that the two direct effects of commitment are zero no longer applies (since players no longer necessarily act selfishly "at the outset").

By Lemma 2, however, a small increase $d\lambda_i > 0$ in, say, i's strategic preference (not necessarily starting from $\lambda_i = 0$), leads to an increase in its own effort of $dX_i^* > 0$. By Lemma 1, j adjusts its effort by $dX_j^* = (-L_i)\, dX_i^* < 0$ in response.

The two strategic effects of an additional commitment are also as before. First, the increase in i's effort yields an increase in the benefits enjoyed by the *other* player of $B_j' dX_i^* > 0$. Second, the induced reduction in j's effort means that i's benefit changes by $B_i'\,(-L_i)\, dX_i^* < 0$.

The direct effect of a small change dX_i^* in i's effort on its own net benefit Π_i, in general, is equal to $(B_i' - C_i')dX_i^*$. Using i's first-order condition from (3), this generalized direct effect can also be written as $(-\lambda_i \theta_i B_j')dX_i^* \leq 0$. Similarly, the direct effect of a small (induced) change dX_j^* on j's net benefit Π_j is equal to $(B_j' - C_j')dX_j^*$. Again, by its first-order condition, the generalized direct effect equals $(-\lambda_j \theta_j B_i')\, dX_j^* = (\lambda_j \theta_j B_i' L_i)\, dX_i^* \geq 0$.

The overall equilibrium impact of an incremental commitment by i on its true objective

[22] Similarly, if marginal costs are identical in the initial equilibrium $C_i'(X_i^*) = C_j'(X_j^*)$, then a small commitment by i always improves global welfare W^*, but has an ambiguous impact on its true objective S_i^*. (In the initial equilibrium, $B_k' = C_k'$ $(k = i, j)$ as both are entirely selfish.)

$S_i = \Pi_i + \theta_i \Pi_j$ takes into account all of these effects, with appropriate weights:

$$\begin{aligned} dS_i^* \;=\; & \underbrace{(-\lambda_i \theta_i B_j') dX_i^*}_{\substack{\text{direct effect} \\ \text{on player } i\ (\leq 0)}} \;+\; \underbrace{\left(-B_i' L_i\right) dX_i^*}_{\substack{\text{strategic effect} \\ \text{on player } i\ (<0)}} \\ & + \underbrace{\theta_i}_{\substack{\text{true altruism} \\ \text{of player } i\ (\in [0,1])}} \times \Big[\underbrace{\left(\lambda_j \theta_j B_i' L_i\right) dX_i^*}_{\substack{\text{direct effect} \\ \text{on player } j\ (\geq 0)}} + \underbrace{\left(B_j'\right) dX_i^*}_{\substack{\text{strategic effect} \\ \text{on player } j\ (>0)}} \Big]. \end{aligned}$$

This decomposition shows that, in general, selfless action reduces a player's own net benefit ($d\Pi_i^* < 0$) but helps the other player ($d\Pi_j^* > 0$). Writing it more compactly yields:[23]

Lemma 4 *The generalized impact of a small unilateral commitment $d\lambda_i > 0$ by player i on her equilibrium true objective satisfies*

$$\frac{dS_i^*}{d\lambda_i} = \left[(1-\lambda_i)\theta_i B_j' - (1-\lambda_j \theta_i \theta_j) B_i' L_i\right] \frac{dX_i^*}{d\lambda_i}.$$

Lemma 4 tells us the marginal equilibrium impact of more altruistic behaviour by player i on her true objective, taking into account its impacts on both i's own contribution effort and the incentive effect on j's contribution. By inspection, it is clear that the impact is ambiguous in general.

Main results. We can now establish the key result that a "full commitment" with $\lambda_i = 1$ is *almost* never optimal for player i.

Proposition 3 *(a) If both players' true preferences are entirely altruistic $\theta_i = \theta_j = 1$, then their optimal commitments $\lambda_i^* = \lambda_j^* = 1$ achieve first-best effort levels;*
(b) If at least one player has partially selfish true preferences $\theta_i < 1$ or $\theta_j < 1$, then optimal commitments $\lambda_i^* < 1$ and $\lambda_j^* < 1$ and both efforts fall short of first-best levels.

Part (a) of the result shows that the first-best outcome is sustainable in our model as long as *both* players want to be entirely unselfish. The intuition is that if both players care about global welfare, neither has an incentive to unilaterally deviate from a full commitment since any such deviation, by construction, causes global welfare to fall.

Part (b) shows that this optimistic conclusion applies only where both players are entirely altruistic. Whenever at least one player places greater weight on domestic welfare in its true objective function, *both* players' optimal commitments fall short of a full com-

[23] The formulae in Proposition 1 ($\lambda_i = \lambda_j = 0$) and Proposition 2 ($\lambda_i = \lambda_j = 0$ and $\theta_i = 1$) can be obtained as special cases of Lemma 4.

mitment, $\lambda_i^* < 1$ and $\lambda_j^* < 1$. In such cases, given the optimal strategic preference chosen at Date 1, player i chooses Date 2 effort to $\max_{X_i} \Omega_i = \Pi_i + \lambda_i^* \theta_i \Pi_j$, with $\lambda_i^* \theta_i < 1$.[24]

Think about the impact of the "last step" towards a full commitment with $\lambda_i = 1$. In this case, the negative direct effect on i is sufficiently negative to entirely offset the *weighted* positive strategic effect on j. The reason is that, with a full commitment, i already internalizes the externality of its choice on j (precisely to the extent it cares about her). Thus the impact of the last step is determined solely by the two remaining effects, the strategic effect on i plus the weighted direct effect on j. This equals $[-(1 - \lambda_j \theta_i \theta_j) B_i' L_i] \, dX_i^* < 0$, and is negative since $\theta_i < 1$ or $\theta_j < 1$ by assumption (and also $\lambda_j \leq 1$, in equilibrium). Therefore, the last step reduces the equilibrium value of i's true objective S_i^*. The same reasoning applies to the other player, so, in equilibrium, $\lambda_i^* < 1$ and $\lambda_j^* < 1$. It is optimal, for instance, that each countries' citizens delegate decision-making regarding public good provision to politicians whose preferences are closer to the national self-interest.

Perhaps the most striking statement of this latter result goes as follows: Suppose i is entirely altruistic, so $\theta_i = 1$, while j is unselfish only to some degree with $\theta_j < 1$. Then part (b) says that the *optimal* commitment by i satisfies $\lambda_i^* < 1$, so a full commitment is dominated by a weaker policy. The optimal way for i to maximize global welfare W is to maximize a strategic objective $\Omega_i = (1 - \lambda_i^*)\Pi_i + \lambda_i^* W$ that is partially skewed towards its own national welfare. In other words, *a player who genuinely wants to maximize global welfare does best by being at least somewhat selfish.*

Intuitively, why can i do better than playing according to its true, entirely altruistic preference? A small decrease in its own effort leads only to a second-order loss in global welfare (by the envelope theorem). But the resulting induced *increase* in the other player's effort creates a first-order gain (whenever the other player is not already choosing the first-best effort). So the reason why full commitment is almost never optimal is what we call "reverse leakage"—a weaker commitment reduces free-riding by the other player.[25]

To further sharpen this argument, we now turn to the opposite limiting case: Our next result shows that, in a range of cases, the optimal commitment for one or both players is a *zero* commitment. More generally, we can show that optimal commitments are often "low", despite players having significantly altruistic preferences.

Proposition 4 *(a) If at least one player's true preference is not entirely altruistic, $\theta_i < 1$ or $\theta_j < 1$, and the ratio of marginal benefits B_i'/B_j' is sufficiently large, then player i's optimal commitment $\lambda_i^* = 0$;*
(b) If players' true preferences $\theta_i > 0$ and $\theta_j > 0$ but both sufficiently small, then players' optimal commitments $\lambda_i^* = \lambda_j^* = 0$.

[24] Proposition 3 thus also rules out any values $\lambda_i^* > 1$ or $\lambda_j^* > 1$ as being sub-optimal. The reason, loosely speaking, is that any such stronger commitment would directly hurt i's own net benefit by more than it can ever strategically benefit j.

[25] A full commitment would become approximately optimal for i in limiting cases where its leakage rate tends to zero. This happens where players' marginal benefits are approximately constant (i.e., $B_k'' \to 0$ for $k = i, j$), or where the other player's production technology is highly inflexible (i.e., $C_j'' \to \infty$) so its effort choice becomes almost non-strategic.

Part (a) of the result essentially gives a non-local version of our earlier findings, from Propositions 1 and 2, that a small commitment by an individual player may not raise S_i^*, or indeed W^*. In extreme cases, it is *optimal* for an entirely altruistic player (when $\theta_i = 1$ but $\theta_j < 1$) to choose her effort level in her own strict self-interest ($\lambda_i^* = 0$).

A further implication is that a policy of zero commitment may welfare-dominate one of full commitment. Suppose that i has a completely altruistic true preference while j is entirely self-interested, $(\theta_i, \theta_j) = (1, 0)$. By Lemma 3, we have that $\lambda_j^* = 0$ irrespective of i's policy. But also, if B_i'/B_j' is sufficiently large, then equilibrium global welfare W^* is higher with zero commitments $(\lambda_i, \lambda_j) = (0, 0)$ than with $(\lambda_i, \lambda_j) = (\ell, 0)$ for *any* $0 \leq \ell \leq 1$ (since then $dW^*/d\lambda_i \leq 0$ for all $\lambda_i \in [0, \ell]$). In this example, a global-welfare oriented country does better by maximizing national welfare than by maximizing global welfare.

The reason for part (b) is that a player who is only somewhat unselfish places too little weight on the positive direct and strategic effects that accrue to the other player for the calculus to overcome the negative impact on its own net benefits. Applying this logic to both players, optimal commitments are zero. Formally, the result requires that altruistic preferences are "sufficiently small", yet a key observation is that this is compatible with large degrees of altruism. To illustrate, let players have identical benefit functions, $B_i(\cdot) = B_j(\cdot)$, with altruism parameters lower than leakage rates, $\theta_k < L_k$ (for $k = i, j$). In this setting, optimal commitments are *zero*, $\lambda_i^* = \lambda_j^* = 0$, even though social preferences could be almost fully altruistic.[26]

A simple corollary is that (sufficiently small) increases in one or both players' true levels of altruism (θ_i and/or θ_j), may, in equilibrium, have no impact at all on the quality of public good provision since they are endogenously offset by crowding-out problems.

Interior commitments. To complete this part of our analysis, and move beyond the limiting cases, we now provide a characterization of players' optimal commitments in an interior equilibrium, in which $(\lambda_i^*, \lambda_j^*) \in (0, 1)^2$ (and thus also $\theta_i > 0$ and $\theta_j > 0$).

Proposition 5 *In an interior equilibrium with $(\lambda_i^*, \lambda_j^*) \in (0, 1)^2$, player i's optimal commitment λ_i^* satisfies*

$$\lambda_i^* = \frac{\left[\theta_i(1 - L_i L_j) - (1 - \theta_i\theta_j)\,(B_i'/B_j')L_i\right]}{\theta_i\,(1 - \theta_i\theta_j L_i L_j)} \in (0, 1),$$

where the equilibrium rates of leakage

$$L_i = \frac{\left[1 + \lambda_j^*\theta_j(B_i''/B_j'')\right]}{\left[1 + (C_j''/|B_j''|) + \lambda_j^*\theta_j(B_i''/B_j'')\right]} \in (0, 1),$$

and player i's equilibrium effort satisfies $X_i^* = C_i'^{-1}\left(B_i' + \lambda_i^*\theta_i B_j'\right) > 0$.

[26] Of course, as long as leakage rates are even higher; this is always true, e.g., where players' marginal costs are approximately constant, $C_k'' \to 0$.

Proposition 5 implicitly describes players' optimal interior commitments, leakage rates, and contribution efforts given their respective benefit and cost functions as well as their true preferences for altruistic behaviour. In principle, a numerical solution for the six unknowns can be obtained by making specific assumptions on the functional forms of $B_k(\cdot)$ and $C_k(\cdot)$. The basic informational requirement is as follows: The ratio of players' marginal benefits, B_i'/B_j', and slopes of marginal benefits, B_i''/B_j''; each player's ratio of the slopes of marginal cost to the slope of marginal benefits, $C_k''/|B_k''|$ (with all functions evaluated at equilibrium); and each player's true preference for unselfishness θ_k $(k = i, j)$

The solution can be simplified under some commonly-made assumptions. Let player k's benefit function $B_k(\cdot) = \mu_k B(\cdot)$, where $\mu_k > 0$ is the weight placed on a *global* benefit function $B(X_i + X_j)$. This has the advantage that the ratios $B_i'/B_j' = B_i''/B_j'' = \mu_i/\mu_j$ become invariant to the details of players' contributions. Also assume that marginal costs and benefits are linear, $B_k'(X_i + X_j) = [\alpha_k - \beta_k (X_i + X_j)]$ and $C_k'(X_k) = \delta_k X_k$, so that $C_k''/|B_k''| = \delta_k/\beta_k$ is constant, too.[27] Optimal commitments can then be determined more easily—as the solution to a system of four equations and four unknowns $(\lambda_i^*, \lambda_j^*, L_i, L_j) \in (0,1)^4$, for given underlying true preferences (θ_i, θ_j).

Inferring how altruistic players are. Suppose it is observed or otherwise estimated that player i's public good contribution appears to be entirely selfish; this corresponds to $\lambda_i^*\theta_i = 0$ in our model. As the above analysis shows, it does *not* follow that this player's underlying true preference is completely selfish. Little or no additional effort can be consistent even with highly altruistic true preferences—simply because it may arise from $\lambda_i^* = 0$ rather than $\theta_i = 0$. So *caution is required in inferring whether or not a player is "being selfish" from her observed behaviour.*

More generally, how does players' optimal altruism compare with true preferences?

Proposition 6 *(a) For true preferences $0 < \theta_i = \theta_j < 1$, optimal commitments may satisfy $\lambda_i^*\theta_i \neq \lambda_j^*\theta_j$;*
(b) For true preferences $0 < \theta_j < \theta_i$, optimal commitments may satisfy $\lambda_j^*\theta_j > \lambda_i^*\theta_i$;
(c) If true preferences $0 < \theta_j < \theta_i$, as well as $B_i' \geq B_j'$, and $L_i \geq L_j$, optimal commitments in an interior equilibrium satisfy $(\lambda_i^*\theta_i - \lambda_j^*\theta_j) < (\theta_i - \theta_j)$.

Part (a) observes that players with identical true preferences toward altruism may have different degrees of optimal altruism. Except in knife-edge cases, this will always occur if they have different benefit and/or cost functions. Part (b) notes that the general relationship between true and strategic preferences is even less clear-cut. A player who cares more about global welfare may, in equilibrium, be the player whose actions are closer to self-interest. In short, player*s with identical true degrees of altruism may behave differently, and a "more altruistic" player may optimally behave less altruistically than*

[27] This latter assumption is essentially equivalent to the classic analysis of Weitzman (1974) on whether price- or quantity-based regulation is socially preferable. It can be seen as a second-order approximation to the unknown shapes of the underlying cost and benefit functions (see also Barrett, 1994).

another player. Part (c) shows that, in an interior equilibrium, there is a tendency for strategic considerations to compress any cross-player differences in altruism: The difference in optimal degrees of altruism is often less than that of true degrees of altruism.

Taken together, these findings pose obvious challenges for making cross-country inferences on true degrees of altruism based on countries' observed choices.

5 Robustness of the main results

The main results from the benchmark model are that the welfare impact of an altruistic commitment is ambiguous (Propositions 1 and 2), a full commitment is optimal only if both players have entirely unselfish true preferences (Proposition 3), and, in some cases, a zero commitment may be optimal despite significantly altruistic preferences (Proposition 4). We have emphasized these limiting cases because, as explained in detail in this section, we believe that these insights are robust to a large variety of changes to the model's specification.

Discussion. In the above, we have, for simplicity, written each player's strategic objective as a weighted average of the form $\Omega_k = (1-\lambda_k)\Pi_k + \lambda_k S_k$. But observe that our results on the ambiguous impact of a small altruistic commitment do not rely on Ω_k at all. Player i, say, raises her level of effort by a small amount dX_i—what exactly induces this is irrelevant for the local results of Propositions 1 and 2. Moreover, our result that a full commitment is almost never optimal, due to reverse leakage, is based on small "profitable" (that is, S_i^*-increasing) deviations away from the case where $\lambda_k = 1$. Again, this analysis does not depend importantly on the functional form of $\Omega_k(\cdot)$.[28] (Of course, the precise values of $(\lambda_i^*, \lambda_j^*)$ in an *interior* equilibrium are, in general, sensitive to the formulation of $\Omega_k(\cdot)$, for instance in Proposition 5.)

Our results are also robust to different definitions of "global welfare". Our above definition $W = \Pi_i + \Pi_j$ is appropriate for the climate problem and corresponds to the usage of the SCC. In some applications, one might instead consider social welfare to be $W = S_i + S_j$, where $S_i = \Pi_i + \theta_i \Pi_j$ (symmetrically for j), which *directly* incorporates players' altruistic preferences. The only results potentially affected are Propositions 2 and 3. It is easy to see, using $W = (1+\theta_j)\Pi_i + (1+\theta_i)\Pi_j$, that Proposition 2 certainly goes through as above if $\theta_i = \theta_j$. More generally, part (a) becomes that $dW^* > 0$ if $B_i' \leq [(1+\theta_i)/(1+\theta_j)]B_j'$, while part (b) remains unchanged. Moreover, Proposition 3 continues to hold, noting only that the first-best effort levels that maximize W will, in general, differ from the benchmark model (again, unless $\theta_i = \theta_j$).

We also assumed that a player's benefits $B_k(X_i + X_j)$ depend on the *unweighted* sum of efforts, which is an appropriate assumption for a range of applications. But observe that the underlying intuition does not depend crucially on the pure public good property.

[28] For instance, we could write $\Omega_k = (1-\lambda_k)\Pi_k^{\phi_k} + \lambda_k S_k^{1-\phi_k}$, with the weight $\phi_k \in (0,1)$, or more generally $\Omega_k = h_k((1-\lambda_k)\Pi_k, \lambda_k S_k)$, where the function $h_k(\cdot,\cdot)$ is strictly increasing in each of its arguments. Key is that maximizing any of these alternative strategic objective boils down to maximizing S_k whenever $\lambda_k = 1$.

The two important features of our setup, in addition to altruistic motives, are (i) that each player would, as such, like the other player to contribute more ($\partial S_i/\partial X_j > 0$), and (ii) the leakage problem that more effort by one player crowds out the other player ($dX_j^*/dX_i < 0$). Our basic insights also apply to many situations with impure public goods, including examples we discussed in the introduction.[29]

In the remainder of this section, we show that our key results are also robust in several other directions, in particular, to the generalization to $n \geq 3$ players, to moderate degrees of cross-country spillovers in costs, and to alternative representations of altruism in the true objective function S_k, including the "warm glow" of Andreoni (1989, 1990). (Appendix B provides detailed proofs.)

Generalization to $n \geq 3$ players. The analysis quickly gets more complex as the number of players increases; each individual player may have a different benefit and cost function, a different true preference for altruism towards other players, and her own leakage rate. Nevertheless, we can exploit the fact that the model with $n \geq 3$ players remains an aggregative game (Corchón, 1994) at Date 2 when players make contribution decisions.[30] The key is that an increase in i's effort now induces *each* of the $n-1$ other players to cut back; in other words, player-specific leakage rates $L_{ij} \equiv [-R_j'(X_i)]$ are positive. But the overall leakage rate $L_i \equiv \sum_{j \neq i} L_{ij} \in (0,1)$ remains less than 100% and so global contributions rise (corresponding to Lemma 2).[31]

Consider a small commitment by player 1 (beginning in a completely selfish world with $\{\lambda_k\}_{k=1}^n = 0$), and, to illustrate, suppose that $n = 3$. The increase in 1's effort directly raises the net benefits of 2 and 3. It also induces 2 to contribute less, which hurts 1 but now also hurts 3. Similarly, reduced effort by 3 hurts 1 and 2. So 1 is hurt twice due to leakage, and it now induces positive and negative effects on each of the other two players—which it cares about depending on its true preference for altruism. In general, the number of effects to take into account is of order n^2.

The welfare impact of a small commitment is, as before, ambiguous. In particular, we can show that our earlier conditions from Proposition 1 generalize cleanly: $dS_i^* > 0$ holds whenever i has a marginal benefit that is (weakly) below average $B_i' \leq \overline{B}_{-i}'$, its true preference exceeds the leakage rate $\theta_i > L_i$, and the covariance between the $n-1$ other players' marginal benefits and their leakage rates is non-negative $cov(B_j', L_{ij}) \geq 0$.[32] The latter condition ensures that those players that cut effort back more strongly are also those

[29] Our assumption that players hold Nash conjectures when choosing effort levels at Date 2 also does not seem critical for our results (that is, we could let i conjecture a non-zero response by j when choosing X_i).

[30] In an aggregative game, each player's payoff depends only on her own action and a summary statistic of all other players' actions (in our case, the unweighted sum of others' efforts).

[31] We note that many of our basic insights would also apply in settings in which *some* player-specific leakage rates are zero (or even negative), so long as the overall leakage rate remains sufficiently high.

[32] Formally, we define this covariance based the following:

$$\frac{1}{(n-1)}\sum_{j\neq i} B_j'(L_i - L_{ij}) = \left(\frac{1}{(n-1)}\sum_{j\neq i} B_j'\right)\left(\frac{1}{(n-1)}\sum_{j\neq i}(L_i - L_{ij})\right) - cov(B_j', L_{ij}).$$

which benefit more strongly from i's altruism. (Signing $cov(B_j', L_{ij})$ is an empirical issue which may have different answers for different public good problems and different players therein.) Conversely, if $B_i'/\overline{B}'_{-i}$ is sufficiently large and $cov(B_j', L_{ij}) \leq 0$, then $dS_i^* < 0$. The conditions for $dW^* \lessgtr 0$ follow similarly, and generalize Proposition 2.

Our "reverse leakage" intuition also applies with $n \geq 3$ players. Say player 1 cares about global welfare and suppose it engages in a full commitment. Since player-specific leakage rates are positive, a small decrease in its effort—which comes at a second-order loss to global welfare—induces each of the $n-1$ other players to increase effort. As long as at least one of the other players was doing too little from a global-welfare viewpoint, the reverse-leakage effect leads to at least one first-order gain. In a sense, a larger number of players makes altruistic behaviour more difficult to justify—*a single "bad apple" is enough to make all players' optimal commitments fall short of first-best.*

Using analogous arguments to those in the benchmark analysis, a zero commitment is optimal for a player who derives sufficiently low marginal benefits or with a sufficiently low (yet non-zero) true degree of altruism.

Cross-country cost spillovers. It is frequently argued, notably in the context of the development of renewable energy sources such as solar and wind, that more CO_2 abatement effort by one country creates knock-on benefits for other countries in that it leads to a reduction in their (marginal) abatement costs, for instance, due to learning-curve effects and technology spillovers.

We can represent such a scenario in the model by considering a more general cost function $C_i(X_i, X_j)$ which depends on both countries' effort levels. We assume that $B_i'' < \partial^2 C_i/\partial X_i \partial X_j < 0$ and $\partial C_i/\partial X_j < 0 \leq \partial^2 C_i/\partial X_j^2$; more investment by country j reduces country i's total cost (at a decreasing rate), and also reduces its marginal cost (but not too strongly, $\partial^2 C_i/\partial X_i \partial X_j > B_i''$). These conditions are sufficient to ensure the model remains well-behaved, and leakage rates remain strictly positive—as is consistent with the existing empirical evidence for climate policy.[33]

Cost spillovers are a two-edged sword. Consider the impact of a small altruistic commitment by i. This increases j's national welfare, appropriately weighted, by $\theta_i(B_j' - \partial C_j/\partial X_i)dX_i^* > 0$, which, loosely speaking, is more positive than before. However, the resulting carbon leakage affects i's own national welfare by $-[(B_i' - \partial C_i/\partial X_j)L_i]dX_i^* < 0$, which is more negative than before. Thus, *cost spillovers make it easier to help the other player, but also exacerbate the leakage problem for the altruistic player.* The welfare impact of a small commitment thus remains ambiguous in general—even if i cares about global welfare (with $\theta_i = 1$).

Our reverse-leakage argument also applies, again in a sense more strongly than in the

[33] If cost spillovers are so strong that they turn leakage rates negative, this would alter the fundamental nature of the public-good game. Then the externality between countries would turn positive at the margin, and a full commitment would generally become optimal (or an even stronger commitment insofar as a country is able to commit to placing less than full weight on its own national welfare). See also our concluding discussion in Section 7.

benchmark model. A small reduction in i's commitment away from the full-commitment level still induces j to increase effort; this is now doubly beneficial in that it increases i's benefits but now also reduces her costs.

In some cases, again, a zero commitment is optimal. Essentially, this happens when the net benefit to i of additional effort by j, $(B_i' - \partial C_i/\partial X_j)$, is sufficiently larger than the net benefit to j of additional effort by i, $(B_j' - \partial C_j/\partial X_i)$. This generalizes the relative marginal benefits condition from Proposition 4(a); a zero commitment is more likely for a player who enjoys relatively strong cost spillovers from others' efforts. It is clear that the conditions from Proposition 4(b) apply here too.

Other altruistic objective functions. Players' true objective functions in our benchmark model represent "pure" altruism: A player directly cares about another player's welfare. Suppose more generally that i's true objective function $S_i = (1 - \widetilde{\theta}_i)\Pi_i + \widetilde{\theta}_i\Psi_i$, where $\Psi_i(X_i, X_j)$. Define $\Phi_i \equiv (\Psi_i - \Pi_i)$ so that we can write $S_i = \Pi_i + \widetilde{\theta}_i\Phi_i$.[34] Policy decisions are delegated, say, by way of a strategic objective $\Omega_i = (1 - \widetilde{\lambda}_i)\Pi_i + \widetilde{\lambda}_i S_i$, where the strategic preference $\widetilde{\lambda}_i \in [0, 1]$.

Various alternative objectives can be represented this way, including forms of "impure" altruism. For example, it has been argued that contributing to a public good yields a "warm glow" (Andreoni 1989, 1990); such objectives are essentially equivalent to $\Phi_i = g_i(X_i)$, so the player derives direct utility-benefits from her effort (with $g_i'(\cdot) > 0$ and $g_i''(\cdot) \leq 0$). Some countries might even be willing to "ignore" some of the CO_2 abatement costs they incur in unilateral climate action, that is, $\Phi_i = C_i(X_i)$ and $\widetilde{\theta}_i \in [0, 1]$ is the true preference for cost understatement. Conversely, some countries may overestimate the benefits of their actions, for example, by using "too high" a discount factor in policy analysis.

Under of any of these objectives, a stronger commitment by a player increases her own contribution but raises global contributions by less (Lemmas 1 and 2). For other objectives, these conclusions hold under mild assumptions; sufficient conditions are $\partial\Phi_i/\partial X_i > 0$ and $\partial\Phi_i/\partial X_j \geq 0$ as well as $[(B_i'' - C_i'') + \partial^2\Phi_i/\partial X_i^2] < 0$ and $\partial^2\Phi_i/\partial X_i\partial X_j \leq 0$.

Consider the impact of a small commitment $dX_i^*/d\widetilde{\lambda}_i > 0$, starting from $\widetilde{\lambda}_i = \widetilde{\lambda}_j = 0$. By similar arguments as in the benchmark model, this increases i's true objective by $\widehat{\theta}_i\,(\partial\Phi_i/\partial X_i)\,(dX_i^*/d\widetilde{\lambda}_i) > 0$. However, crowding-out affects i's true objective according to $-\{[B_i' + \widetilde{\theta}_i\,(\partial\Phi_i/\partial X_j)]L_i\}(dX_i^*/d\widetilde{\lambda}_i) < 0$. The sign of the overall welfare impact $(dS_i^*/d\widetilde{\lambda}_i)_{\widetilde{\lambda}_i=\widetilde{\lambda}_j=0}$ is thus ambiguous in general, even if $\widetilde{\theta}_i = 1$. It is also not difficult to confirm that the impact of a small commitment $d\widetilde{\lambda}_i > 0$ on equilibrium global welfare W^* is *exactly* as in Proposition 2 above.

Our reverse-leakage argument applies since i always wants j to increase its effort (that is, $\partial S_i/\partial X_j > 0$). A small reduction away from a full commitment (with $\widetilde{\lambda}_i = 1$) leads to a first-order increase in S_i^*, such that the optimal commitment $\widetilde{\lambda}_i^* < 1$. This is *always* the case for any of the "impure" forms of altruism discussed above, including the "warm

[34] In the benchmark model, $\Psi_k = W$ (for $k = i, j$) and so $\Phi_i = \Pi_j$ (and vice versa).

glow" (even if j is already choosing its effort to maximize global welfare).[35]

Finally, under some conditions on $\widehat{\theta}_i$ and $\Phi_i(\cdot)$, a zero commitment becomes optimal; it turns out that a simple sufficient condition for $\widetilde{\lambda}_i^* = 0$ is $\partial\Phi_i/\partial X_i \leq B_i'$, that is, a higher national contribution raises benefits no less than the altruistic part of i's objective.

6 Further properties of the model

To close our analysis, we highlight two other features of the benchmark model: First, the impact of altruistic behaviour on leakage at Date 2, and, second, the strategic properties of players' commitments at Date 1 of the game.

The impact of altruism on leakage. It may seem natural to conjecture that altruistic behaviour tends to mitigate free-riding and reduce leakage. It turns out that, under some circumstances, this intuition is quite misleading:

Proposition 7 *(a) Suppose that $B_j''' \leq 0$ and $C_j''' \leq 0$. Player i's leakage rate is higher when player j also has an unselfish commitment than when player j acts entirely selfishly, $L_i|_{\lambda_j^*>0} > L_i|_{\lambda_j^*=0}$.*
(b) Suppose that B_i''/B_j'' and B_j''/C_j'' are both constant. Player i's leakage rate increases in player j's commitment, $(dL_i/d\lambda_j)_{\lambda_j^*\geq 0} > 0$.

To understand the result, recall j's first-order condition for its effort choice, $\partial\Omega_j/\partial X_j = (\partial\Pi_j/\partial X_j) + \lambda_j\theta_j\,(\partial\Pi_i/\partial X_j) = 0$. The overall rate of leakage can hence be thought of in two parts: Firstly, a selfish component $\partial\Pi_j/\partial X_j$, and, secondly, an altruistic component $\partial\Pi_i/\partial X_j$ (which, in equilibrium, receives weight $\lambda_j^*\theta_j$). The key point is that the altruistic component has a leakage rate of 100%. To see why, observe that holding $\partial\Pi_i/\partial X_j = B_i'(X_i + X_j)$ fixed (along j's reaction function) in response to a small increase in i's effort $dX_i > 0$ requires a decrease in j's effort $dX_j = -dX_i < 0$ that is exactly offsetting. Greater weight on the altruistic part therefore certainly tends to increase the overall leakage rate as long as the selfish part does not decline as a result. The conditions given in Proposition 7 are, respectively, (a) sufficient for the selfish part to not decline, and (b) necessary and sufficient for it to stay constant.[36]

Although global welfare may (but need not) be higher when players pursue altruistic objectives, the associated leakage rates can also be higher than with self-interested behaviour. Put differently, *altruism can worsen the free-riding problem at the margin*. From an empirical point of view, this suggests a surprising possibility: Rates of carbon leakage associated with unilateral climate action may be "high" precisely *because* countries are

[35] Compared to the benchmark model, the disadvantage of such other objective functions is that the global-welfare preference ("social cost of carbon") is no longer a natural special case.

[36] We note that the results of Proposition 7 would also apply if players' altruistic commitments were not chosen optimally; they speak generally to the impact of altruism on leakage, not necessarily that of optimal altruism on leakage.

behaving altruistically. The more general point is that leakage rates—though a useful and important statistic—are not always a reliable welfare indicator.

Policy commitments: Strategic substitutes or complements? Finally, we explore the strategic properties of the game at Date 1 where players choose their respective policy commitments. For this analysis, it will be useful to define a "leakage-commitment" elasticity:

$$\eta_{ij} \equiv \frac{dL_i/L_i}{d\lambda_j/\lambda_j}.$$

This measures the elasticity of player i's leakage rate with respect to a stronger commitment by player j. For example, under the conditions of Proposition 7(b), we found $(dL_i/d\lambda_j)_{\lambda_j^* \geq 0} > 0$, implying that, in such cases, the leakage-commitment elasticity $\eta_{ij} > 0$ (in an interior equilibrium). Using this metric, we obtain the following result:

Proposition 8 *Consider an interior equilibrium with* $(\lambda_i^*, \lambda_j^*) \in (0,1)^2$.
(a) Suppose that B_i'/B_j' *is constant. Player* i*'s optimal commitment varies with player* j*'s commitment according to*

$$sign\left\{\frac{d\lambda_i^*}{d\lambda_j}\right\} = sign\left\{\frac{1}{\eta_{ij}} - \left(\frac{1-\lambda_j^*\theta_i\theta_j}{\lambda_j^*\theta_i\theta_j}\right)\right\}$$

where $\eta_{ij} \equiv [(dL_i/L_i)/(d\lambda_j/\lambda_j)]_{\lambda_k=\lambda_k^*}$ *is the (equilibrium) leakage-commitment elasticity.*
(b) Suppose that B_i'/B_j', B_j''/C_j'' *and* B_i''/B_j'' *are all constant. Then the leakage-commitment elasticity* $\eta_{ij} \in (0,1)$*, and so* $d\lambda_i^*/d\lambda_j > 0$ *if* $\lambda_j^*\theta_i\theta_j \geq \frac{1}{2}$ *while* $d\lambda_i^*/d\lambda_j < 0$ *if* $\lambda_j^*\theta_i\theta_j$ *is sufficiently small.*

In general, therefore, it is ambiguous whether players' policy commitments are strategic substitutes or strategic complements. If the leakage-commitment elasticity $\eta_{ij} > 0$ and the "joint-altruism" term $\lambda_j^*\theta_i\theta_j$ is sufficiently small, then we have strategic substitutes $(d\lambda_i^*/d\lambda_j < 0)$. By contrast, if $\eta_{ij} \leq 0$ *or* if $\lambda_j^*\theta_i\theta_j$ is sufficiently large, then we have strategic complements $(d\lambda_i^*/d\lambda_j > 0)$. Of course, the results from Proposition 7 suggests that, in many cases, $\eta_{ij} > 0$; if so, the level of $\lambda_j^*\theta_i\theta_j$ becomes the main determinant of a commitment's strategic properties.[37]

Proposition 8 strikes us as interesting for several reasons. First, the strategic properties of the game in commitment space (Date 1) may thus differ from those of the effort game (Date 2)—which is always characterized by strategic substitutes (Lemma 1). Second, it has a similar flavour to an intuition found in other public good models: *Commitments are strategic complements if they are already "high", but strategic substitutes when they are "low"* (see, e.g., the tipping-point analysis of Heal and Kunreuther, 2010).

[37] Cases with $\eta_{ij} < 0$ seem relatively unusual but could occur where C_j''' is positive and large near equilibrium, in other words, where effort costs are highly convex.

We have not been able to derive a full set of results on the impact of a sequential move order at Date 1, but conjecture that, in general, it is ambiguous whether sequential commitment makes a difference, and if any such difference raises or reduces welfare.[38]

Illustrative example. We present a simplified example to illustrate our results from this section, and to link them to our earlier findings. Suppose that players are symmetric with identical benefit and cost functions, but asymmetric in that their true levels of altruism differ, where $\theta_i = 1$ but $\theta_j < 1$. In particular, assume $B_k'(X_i + X_j) = \alpha - X_i - X_j$, so that $B_i'/B_j' = B_i''/B_j'' = 1$ is constant, and that $C_k''/|B_k''| = 1$ is constant (for $k = i, j$).

Without any further calculations, we know that $\lambda_i^* > 0$ by Proposition 1(a) but also that $\lambda_i^* < 1$ and $\lambda_j^* < 1$ by Proposition 3(b); here, we consider two scenarios:

First, suppose, that θ_j is sufficiently small, in particular $\theta_j \gtrsim \frac{6}{13}$ that $\lambda_j^* \gtrsim 0$. Using Lemma 4, i's optimal commitment then solves the first-order condition $dS_i^*/d\lambda_i \simeq (1 - \lambda_i) - L_i(\lambda_j) = 0 \Longrightarrow \lambda_i^* \simeq \left[1 - L_i(\lambda_j^*)\right] \in (0, 1)$, where the equilibrium leakage rate $L_i(\lambda_j^*) = (1 + \lambda_j^*\theta_j)/\left(2 + \lambda_j^*\theta_j\right)$. It is then immediate that $d\lambda_i^*/d\lambda_j \simeq -(dL_i/d\lambda_j)_{\lambda_j = \lambda_j^*} < 0$. Put differently, in this example, Proposition 7(b) and the strategic substitutes part of Proposition 8(b) are two sides of the same coin: i regards j's commitment as a strategic substitute because a stronger commitment by j drives up the leakage rate associated with i's own policy.[39] In summary, $(\theta_i, \theta_j) = (1, \frac{6}{13}) \Longrightarrow (\lambda_i^*, \lambda_j^*) = (\frac{1}{2}, 0)$, which also confirms one of main arguments, based on Proposition 4, that large degrees of true altruism may only lead to much lower degrees of optimal/observed altruism.

Second, and by contrast, assume that θ_j is sufficiently large such that $\lambda_j^*\theta_j \geq \frac{1}{2}$.[40] Then we have that $d\lambda_i^*/d\lambda_j > 0$ by the other part of Proposition 7(b), and so commitments become strategic complements. In such cases, loosely speaking, the additional commitment by j is sufficiently valuable, in that it raises the marginal return on i's own commitment, to offset the adverse impact of the higher leakage rate.

[38] That is, what happens if one player is a leader in choosing her strategic preference? A partial analysis goes as follows: Whenever policy commitments are strategic complements (see Proposition 8), a stronger commitment by the first-mover induces the follower to also raise her effort and so global contributions at Date 2 rise. This, however, does not characterize optimal sequential commitments. A player's optimal commitment in our benchmark model may be zero (Proposition 4); in such cases, a change to sequential moves may make no difference insofar as a player remains "stuck in a corner" (at $\lambda_j^* = 0$). Moreover, while global contributions may rise, the preceding analysis highlights that this need not yield a welfare improvement. Finally, we have not characterized the conditions under which a player would, in fact, want to become a first-mover in the first place. Nonetheless, based on these arguments, we arrive at the conjecture described in the main text.

[39] To check when indeed $\lambda_j^* \gtrsim 0$, use j's first-order condition to obtain $dS_j^*/d\lambda_j = (1 - \lambda_j)\theta_j - (1 - \lambda_i\theta_j)L_j = 0 \Longrightarrow (1 - \lambda_j^*) = \left(\theta_j^{-1} - \lambda_i^*\right) L_j$. It follows that $\lambda_j^* \gtrsim 0$ whenever $(\theta_j^{-1} - \lambda_i^*)L_j \lesssim 1 \Longleftrightarrow \theta_j \gtrsim (L_j^{-1} + \lambda_i^*)^{-1}$. Now using $\lambda_i^* \simeq (1 - L_i)$ as well as $L_j = (1 + \lambda_i^*)/\left(2 + \lambda_i^*\right) \simeq (2 - L_i)/(3 - L_i)$, this can also be written as $\lambda_j^* \gtrsim 0$ whenever $\theta_j \gtrsim \left(\frac{(3 - L_i)}{(2 - L_i)} + (1 - L_i)\right)^{-1} \simeq \frac{6}{13}$ since then $L_i = (1 + \lambda_j^*\theta_j)/(2 + \lambda_j^*\theta_j) \gtrsim \frac{1}{2}$.

[40] By Proposition 3(a), we have that $\lambda_j^* \to 1$ as $\theta_j \to 1$, so also $\lambda_j^*\theta_j \to 1$. Then, by continuity, we also have that $\lambda_j^*\theta_j \geq \frac{1}{2}$ for θ_j sufficiently large.

7 Concluding remarks

We have studied the welfare impact of altruism in a model of public good provision, and introduced a notion of "optimal altruism". Altruistically-minded—yet rational—players take into account the incentive effects of their actions on other players. Due to crowding-out effects, optimal altruistic commitments are *almost* always weaker than the true willingness to pursue unselfish action—and, in a range of cases, *much* weaker. Thereby, we have highlighted that players who derive an above-average marginal benefit from contributions, as well as those facing a high leakage rate, will find it more difficult to follow through on their altruistic preferences—and that altruistic behaviour may actually intensify crowding-out at the margin. We have argued that our main results—which mostly emphasize limiting cases—are robust to a variety of natural changes in model specification, including different types of public good problems and different representations of purely and impurely altruistic preferences.

We can relate our findings to the unilateral climate-policy initiatives discussed in the introduction. By incorporating countries' social preferences we can, in principle, explain any outcome between the standard self-interested equilibrium and first-best. So the unilateral actions observed at the local, national, and regional levels *might* indeed be driven by altruistic preferences.

Our equilibrium analysis yields some sharper conclusions. Under our assumptions, it is not optimal for an individual country—or any subset of countries—to unilaterally commit to taking the full "social cost of carbon" into account in domestic policy. A weaker commitment is better because of reverse-leakage: Others are induced to do more, and this is socially (more) valuable. By contrast, using the SCC only in a selected range of projects may seem broadly consistent with our results.

We can also provide a rationale for a puzzle: Little or no action beyond "business-as-usual" by apparently altruistically-minded players. In our model, social preferences are necessary—but not sufficient—for countries to deviate from their self-interested levels of contribution. Even large degrees of altruism (or increases in altruism) can, in equilibrium, be negated by leakage, so optimal commitments are zero. We have highlighted that it will generally be difficult to infer a player's true preferences from its observed public good contribution.

Our model provides what seems a natural way of thinking about the role that altruism can play in public good problems characterized by the absence of central mechanism designer. The basic message from our analysis is somewhat pessimistic: The tension between altruism and crowding-out effects makes it more difficult to improve public good provision. What, then, could lead to more favourable outcomes?

First, our model has examined the impact of altruism in a non-cooperative setting. In practice, unilateral climate action, for example, by the EU and others has taken place "in the shadow" of evolving cooperative talks between countries, and there may be strategic interaction between them. One view is that leadership in form of unilateral initiatives

signals a willingness to cooperate and thus facilitates agreement; some recent papers, however, have noted that unilateral policy may, in fact, undermine future negotiations (Beccherle and Tirole 2011; Harstad 2012).

Second, and perhaps most obviously, leakage rates may be zero or even negative in some situations, i.e., public good contributions are strategic complements. Then players may find it optimal to follow through on their altruistic preferences—and perhaps try to find ways to commit to doing even more. However, as noted above, negative leakage is not a particularly common feature of public good models, and we are not aware of any such empirical evidence for our application to climate policy.

Third, and related, the relevant contracting space may be richer. Players might be able to make their commitments conditional on what other players are doing. For example, the EU has previously considered a commitment to augment its 2020 carbon-emissions reduction from 20% to 30% if less-developed countries agree to certain abatement targets—although this commitment was never actually activated.[41] In terms of our model, this can be interpreted as an attempt to "change the game": i's conditional commitment turns j's leakage rate negative, at least over some range, and thus encourages it to do more. (But note that i's own leakage rate remains positive.) Again, the existing literature draws mixed conclusions. While conditional commitment is superior to unconditional commitment in some situations (Hoel 1991), matching contributions are ineffective in others (Boadway, Song and Tremblay 2007) and may actually worsen public good provision if commitments are only made by a subset of players (Buchholz, Cornes and Rübbelke 2012).[42]

The existing public goods literature on the role of conditional commitments and the evolution of policy negotiations over time has, however, paid little attention to the role of altruistic preferences. Combining the optimal-altruism approach presented in this paper—different players are altruistic to different degrees and recognize the incentive effects of their actions—with such richer contracting environments may be an interesting and important topic for future research.

References

Andreoni, James (1989). Giving With Impure Altruism: Applications to Charity and Ricardian Equivalence. *Journal of Political Economy* 97, 1447–1458.

Andreoni, James (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *Economic Journal* 100, 464–477.

[41] Perhaps the most well-known example of such conditional commitments comes from charitable giving, where an initial large donor promises to match contributions by subsequent donors according to some agreed rule. It is interesting how such matching contributions play very different roles across different kinds of public good problems.

[42] Better outcomes can—at least in some situations—also be achieved by different forms of "side contracting", for example, if players can make side payments contingent on other players' actions (Jackson and Wilkie 2005) or if different policy domains—such as environmental policy and trade policy—can be linked (Harstad 2010).

Babiker, Mustafa H. (2005). Climate Change Policy, Market Structure and Carbon Leakage. *Journal of International Economics* 65, 421–445.

Barrett, Scott (1994). Self-Enforcing International Environmental Agreements. *Oxford Economic Papers* 46, 878–894.

Barrett, Scott (2005). The Theory of International Environmental Agreements. In: Karl-Göran Mäler and Jeffrey R. Vincent (eds.), *Handbook of Environmental Economics*, Volume 3, Elsevier.

Beccherle, Julien and Jean Tirole (2011). Regional Initiatives and the Cost of Delaying Binding Climate Change Agreements. *Journal of Public Economics* 95, 1339–1348.

Becker, Gary S. (1974). A Theory of Social Interactions. *Journal of Political Economy* 82, 1063–1094.

Bergstrom, Theodore C. (1989). A Fresh Look at the Rotten Kid Theorem—and Other Household Mysteries. *Journal of Political Economy* 97, 1138–1159.

Bergstrom, Theodore, Lawrence Blume and Hal Varian (1986). On the Private Provision of Public Goods. *Journal of Public Economics* 29, 25–49.

Boadway, Robin, Zhen Song and Jean-François Tremblay (2007). Commitment and Matching Contributions to Public Goods. *Journal of Public Economics* 91, 1664–1683.

Bolton, Gary E. and Axel Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review* 90, 166–193.

Buchholz, Wolfgang, Richard Cornes and Dirk Rübbelke (2012). Potentially Harmful International Cooperation on Global Public Good Provision. Working Paper at University of Regensburg, July.

Copeland, Brian R. and M. Scott Taylor (2005). Free Trade and Global Warming: A Trade Theory View of the Kyoto Protocol. *Journal of Environmental Economics and Management* 49, 205–234.

Corchón, Luis C. (1994). Comparative Statics for Aggregative Games: The Strong Concavity Case. *Mathematical Social Sciences* 28, 151–165.

Cornes, Richard and Roger Hartley (2007). Aggregative Public Good Games. *Journal of Public Economic Theory* 9, 201–219.

Cornes, Richard and Todd Sandler (1996). *The Theory of Externalities, Public Goods and Club Goods*. Cambridge University Press.

DECC (2009). Climate Change Act 2008: Impact Assessment. Department of Energy and Climate Change (London, United Kingdom), March.

Edgeworth, Francis Y. (1881). *Mathematical Physics: An Essay on the Application of Mathematics to the Moral Sciences*. Kegan Paul.

EPA (2014). Regulatory Impact Analysis for the Proposed Carbon Pollution Guidelines for Existing Power Plants and Emission Standards for Modified and Reconstructed Power Plants. Environmental Protection Agency (Washington DC, United States), June.

Fehr, Ernst and Klaus M. Schmidt (1999). A Theory Of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics* 114, 817–868.

Greenstone, Michael, Elizabeth Kopits and Ann Wolverton (2013). Developing the Social Cost of Carbon for U.S. Regulatory Analysis: A Methodology and Interpretation. *Review of Environmental Economics and Policy* 7, 23–46.

Harstad, Bård (2010). Do Side Payments Help? Collective Decisions and Strategic Delegation. *Journal of the European Economic Association* 6, 468–477.

Harstad, Bård (2012). The Dynamics of Climate Agreements. Working Paper at Northwestern University, August.

Heal, Geoffrey and Howard Kunreuther (2010). Social Reinforcement: Cascades, Entrapment, and Tipping. *American Economic Journal: Microeconomics* 2, 86–99.

Heifetz, Aviad, Chris Shannon and Yossi Spiegel (2007). What to Maximize if You Must. *Journal of Economic Theory* 133, 31–57.

Hoel, Michael (1991). Global Environmental Problems: The Effects of Unilateral Actions Taken by One Country. *Journal of Environmental Economics and Management* 20, 55–70.

IPCC (2007). Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Chapter 11: Mitigation from a Cross-Sectoral Perspective. Cambridge University Press.

Jackson, Matthew O. and Simon Wilkie (2005). Endogenous Games and Mechanisms: Side Payments Among Players. *Review of Economic Studies* 72, 543–566.

Kosfeld, Michael, Akira Okada and Arno Riedl (2009). Institution Formation in Public Goods Games. *American Economic Review* 99, 1335–1355.

Lange, Andreas and Carsten Vogt (2003). Cooperation in International Environmental Negotiations Due to a Preference for Equity. *Journal of Public Economics* 87, 2049–2067.

Levrahi, David and Leonard J. Mirman (1980). The Great Fish War: An Example Using a Dynamic Cournot-Nash Solution. *RAND Journal of Economics* 11, 322–334.

Olson, Mancur and Richard Zeckhauser (1966). An Economic Theory of Alliances. *Review of Economics and Statistics* 48, 266–279.

Persson, Torsten and Guido Tabellini (1993). Designing Institutions for Monetary Stability. *Carnegie-Rochester Conference Series on Public Policy* 39, 53–84.

Rabin, Matthew (1993). Incorporating Fairness Into Game Theory and Economics. *American Economic Review* 83, 1281–1302.

Ritz, Robert A. (2009). Carbon Leakage under Incomplete Environmental Regulation: An Industry-Level Approach. Department of Economics, Oxford University, November.

Roelfsema, Hein (2007). Strategic Delegation of Environmental Policy Making. *Journal of Environmental Economics and Management* 53, 270–275.

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Harvard University Press.

Segendorff, Björn (1998). Delegation and Threat in Bargaining. *Games and Economic Behavior* 23, 266–283.

Sobel, Joel (2005). Interdependent Preferences and Reciprocity. *Journal of Economic Literature* 43, 392–436.

Stavins, Robert N. (2011). The Problem of the Commons: Still Unsettled after 100 Years. *American Economic Review* 101, 81–108.

Stern, Nicholas (2008). The Economics of Climate Change. *American Economic Review* 98, 1–37.

Sunstein, Cass R. (2007). Of Montreal and Kyoto: A Tale of Two Protocols. *Harvard Environmental Law Review* 31, 1–65.

Tol, Richard S. J. (2012). A Cost-Benefit Analysis of the EU 20/20/2020 Package. *Energy Policy* 49, 288–295.

Vickers, John (1985). Delegation and the Theory of the Firm. *Economic Journal* 95, 138–147.

Watkiss, Paul and Chris Hope (2011). Using the Social Cost of Carbon in Regulatory Deliberations. *WIREs Climate Change* 2:6, 886–901.

Weitzman, Martin L. (1974). Prices vs Quantities. *Review of Economic Studies* 41, 477–491.

Appendix A: Proofs

Proof of Lemma 2. Observe that for player i, say, $dX_i^*/d\lambda_i = \partial X_i^*/\partial\lambda_i + R_i'(X_j^*)[dX_j^*/d\lambda_i]$ and $dX_j^*/d\lambda_i = R_j'(X_i^*)[dX_i^*/d\lambda_i]$, so that

$$\frac{dX_i^*}{d\lambda_i} = \frac{\partial X_i^*/\partial\lambda_i}{\left[1 - R_i'(X_j^*)R_j'(X_i^*)\right]} = \frac{\partial X_i^*/\partial\lambda_i}{(1 - L_iL_j)}.$$

The denominator of this expression is positive by Lemma 1. Differentiating i's first-order condition from (3) yields that the numerator $\partial X_i^*/\partial\lambda_i = \theta_i B_j'/(-B_i'' + C_i'' - \lambda_i\theta_i B_j'')$, from which the result is immediate (since $B_k' > 0 > B_k''$ and $C_k'' > 0$, $k = i, j$).

Proof of Lemma 3. Let X_i^s denote the level of effort that solves i's first-order condition $\partial\Pi_i/\partial X_i = 0$ at Date 2. Committing at Date 1 to deviate from this affects i's equilibrium payoff according to

$$\frac{d\Pi_i^*}{dX_i} = \frac{\partial\Pi_i^*}{\partial X_i} + \frac{\partial\Pi_i^*}{\partial X_j}R_j'.$$

The first term, $\partial\Pi_i^*/\partial X_i$, is non-positive: By definition, it equals zero at X_i^s, and it is negative by the concavity of the payoff function Π_i for any $X_i > X_i^s$. The second term, $(\partial\Pi_i^*/\partial X_j)\,R_j'$, is negative since $\partial\Pi_i^*/\partial X_j = B_i' > 0$ and $R_j' \equiv -L_i < 0$ by Lemma 1. So a player k cannot do any better than choosing to her effort to $\max_{X_k}\Pi_k$, as claimed.

Proof of Proposition 3. Before turning to the two parts of the proposition, we first establish that $\lambda_k^* > 1$ cannot be optimal. To see this, note using the formula from Lemma 4 that

$$\frac{dS_i^*}{d\lambda_i} \leq \left[(1-\lambda_i)\theta_i B_j' - (1-\theta_i)B_i'L_i\right]\frac{dX_i^*}{d\lambda_i} \tag{5}$$

since $\lambda_j\theta_j \leq 1$ by our assumption. It follows that $(dS_i^*/d\lambda_i)|_{\lambda_i>1} < 0$ for any $\theta_i \in [0,1]$ such that the optimal commitment must satisfy $\lambda_i^* \leq 1$ (and analogously for j).
For part (a), setting $\theta_i = \theta_j = 1$ in the formula from Lemma 4 shows that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\theta_i=\theta_j=1} = \frac{dW^*}{d\lambda_i} = \left[\left((1-\lambda_i)B_j' - (1-\lambda_j)B_i'L_i\right)\frac{dX_i^*}{d\lambda_i}\right]_{\theta_i=\theta_j=1}.$$

So if j is playing $\lambda_j = 1$, then $(dW^*/d\lambda_i)|_{\lambda_j=1} = [(1-\lambda_i)B_j'](dX_i^*/d\lambda_i) \geq 0$ for all $\lambda_i \in [0,1]$. So i's best response is to also play $\widehat{\lambda}_i(1) = 1$, and so optimal commitments $\lambda_i^* = \lambda_j^* = 1$.
For part (b), we proceed in two steps, looking first at the case where $\theta_i < 1$ and then at the case where $\theta_i = 1$. Suppose that $\theta_i < 1$; then we have using (5) that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_i=1} \leq -\left[(1-\theta_i)B_i'L_i\right]\frac{dX_i^*}{d\lambda_i},$$

and so $(dS_i^*/d\lambda_i)|_{\lambda_i=1} < 0$ whenever $\theta_i < 1$. We conclude that if $\theta_i < 1$, then $\lambda_i^* < 1$. Suppose now that $\theta_i = 1$ but that $\theta_j < 1$. From our previous argument, we know that $\lambda_j^* < 1$, so also $\lambda_j^*\theta_j < 1$. Again using the formula from Lemma 4, we thus have

$$\frac{dS_i^*}{d\lambda_i} < (1-\lambda_i)B_j'\frac{dX_i^*}{d\lambda_i},$$

and so $(dS_i^*/d\lambda_i)|_{\lambda_i=1} < 0$, again implying that $\lambda_i^* < 1$. These arguments establish that if either $\theta_i < 1$ or $\theta_j < 1$, then $\lambda_i^* < 1$ and $\lambda_j^* < 1$.

Proof of Proposition 4. For part (a), use the formula from Lemma 4 to obtain

$$\frac{dS_i^*}{d\lambda_i} = B_j' \left((1-\lambda_i)\theta_i - (1-\lambda_j\theta_i\theta_j)\frac{B_i'}{B_j'}L_i \right) \frac{dX_i^*}{d\lambda_i}.$$

If $\theta_i < 1$ or $\theta_j < 1$, so also $\lambda_j^*\theta_j < 1$ by Proposition 3(b), then $dS_i^*/d\lambda_i < 0$ for all $\lambda_i \in [0, \theta_i^{-1}]$ and $\lambda_j \in [0, \theta_j^{-1}]$ if B_i'/B_j' is sufficiently large, so that the optimal commitment $\lambda_i^* = 0$.
For part (b), observe similarly that $dS_i^*/d\lambda_i < 0$ for all $\lambda_i \in [0, \theta_i^{-1}]$ and $\lambda_j \in [0, \theta_j^{-1}]$ if θ_i is sufficiently small. So if θ_i and θ_j are both sufficiently small, then optimal commitments $\lambda_i^* = \lambda_j^* = 0$ as claimed.

Proof of Proposition 5. In an interior equilibrium, i's best response $\widehat{\lambda}_i(\lambda_j)$ satisfies the first-order condition $dS_i^*/d\lambda_i = 0$. Using the formula from Lemma 4, we thus have

$$\widehat{\lambda}_i\theta_i = \theta_i - (1-\lambda_j\theta_i\theta_j)\frac{B_i'}{B_j'}L_i.$$

Now using this together with the analogous expression for j's best response $\widehat{\lambda}_j(\lambda_i)$ yields that i's optimal commitment $\lambda_i^* \equiv \widehat{\lambda}_i(\widehat{\lambda}_j)$ solves

$$\lambda_i^*\theta_i = \theta_i - \left[(1-\theta_i\theta_j)\frac{B_i'}{B_j'}L_i + \theta_i(1-\lambda_i^*\theta_i\theta_j)L_iL_j \right],$$

which can be arranged to give

$$\lambda_i^* = \frac{\left[\theta_i(1-L_iL_j) - (1-\theta_i\theta_j)\,(B_i'/B_j')L_i\right]}{\theta_i\,(1-\theta_i\theta_jL_iL_j)}$$

as claimed. The expression for the rate of leakage L_i is obtained from Lemma 1 and some rearranging of (4). The expression for country i's effort $X_i^* > 0$ is obtained by rewriting its first-order condition from (3) and noting that the inverse $C_i'^{-1}(\cdot)$ is well-defined under the maintained assumptions $C_i'(\cdot) > 0$, $C_i''(\cdot) > 0$, and $C_i(0) = C_i'(0) = 0$.

Proof of Proposition 6. For part (a), note that this requires an interior equilibrium $(\lambda_i^*, \lambda_j^*) \in (0,1)^2$, and so rewrite the expression for λ_i^* from Proposition 5 as

$$\lambda_i^*\theta_i = \theta_i - \frac{(1-\theta_i\theta_j)}{(1-\theta_i\theta_jL_iL_j)}\left(\theta_iL_j + \frac{B_i'}{B_j'}\right)L_i.$$

Thus the difference in the two players' degrees of altruism satisfies

$$\left(\theta_i\lambda_i^* - \theta_j\lambda_j^*\right) = (\theta_i-\theta_j) - \frac{(1-\theta_i\theta_j)}{(1-\theta_i\theta_jL_iL_j)}\left[(\theta_i-\theta_j)\,L_iL_j + \left(\frac{B_i'}{B_j'}L_i - \frac{B_j'}{B_i'}L_j\right)\right]. \quad (6)$$

Setting $0 < \theta_i = \theta_j < 1$, it is clear that optimal commitments may satisfy $\lambda_i^* \theta_i \neq \lambda_j^* \theta_j$, as claimed; in particular, this occurs whenever $L_i/L_j \neq (B_i'/B_j')^2$. For part (b), since $\theta_j < 1$, it follows that $\lambda_i^* = 0$ by Proposition 4(a) for sufficiently large B_i'/B_j', and so $\lambda_i^* \theta_i = 0$. Using the result from Lemma 4, $dS_j^*/d\lambda_j \geq 0$ for sufficiently small values of $\lambda_j \in [0,1]$ if B_i'/B_j' is sufficiently large, so $\lambda_j^* > 0$, and so $\lambda_j^* \theta_j > 0$. But since $0 < \theta_j < \theta_i$, optimal commitments in this example satisfy $\lambda_j^* \theta_j > \lambda_i^* \theta_i$, as claimed.
For part (c), with $0 < \theta_j < \theta_i$, direct inspection of (6) shows that optimal commitments in an interior equilibrium satisfy $(\lambda_i^* \theta_i - \lambda_j^* \theta_j) < (\theta_i - \theta_j)$ if $B_i' \geq B_j'$ and $L_i \geq L_j$, as claimed.

Proof of Proposition 7. For part (a), it follows from Proposition 4(b) that $\lambda_j^* > 0$ must imply that $\theta_j > 0$, and so also $\lambda_j^* \theta_j > 0$. By contrast, $\lambda_j^* = 0$ also means that $\lambda_j^* \theta_j = 0$. By (4) and Lemma 1, the leakage rate when $\lambda_j^* = 0$ equals

$$L_i|_{\lambda_j^*=0} = \frac{1}{\left(1 + \left[C_j''/(-B_j'')\right]_{\lambda_j^*=0}\right)},$$

while leakage with $\lambda_j^* > 0$ is given by

$$L_i|_{\lambda_j^*>0} = \frac{\left(1 + \lambda_j^* \theta_j \left[B_i''/B_j''\right]_{\lambda_j^*>0}\right)}{\left(1 + \left[C_j''/(-B_j'')\right]_{\lambda_j^*>0} + \lambda_j^* \theta_j \left[B_i''/B_j''\right]_{\lambda_j^*>0}\right)}.$$

So $\left[C_j''/(-B_j'')\right]_{\lambda_j^*>0} \leq \left[C_j''/(-B_j'')\right]_{\lambda_j^*=0}$ is a sufficient condition for $L_i|_{\lambda_j^*>0} > L_i|_{\lambda_j^*=0}$. Furthermore, note that

$$\frac{d}{d\lambda_j}\left[\frac{C_j''}{-B_j''}\right] = \frac{1}{(-B_j'')^2}\left[C_j''' \frac{dX_j^*}{d\lambda_j^*}(-B_j'') - (-B_j''')(1 - L_j)\frac{dX_j^*}{d\lambda_j} C_j''\right]$$

and so (since $dX_j^*/d\lambda_j > 0$ by Lemma 2) we have that

$$\text{sign}\left(\frac{d}{d\lambda_j}\left[\frac{C_j''}{-B_j''}\right]\right) = \text{sign}\left(C_j'''(-B_j'') + B_j'''(1 - L_j)C_j''\right),$$

where the right-hand side is certainly non-positive if $B_j''' \leq 0$ and $C_j''' \leq 0$ (since $L_j \in (0,1)$ by Lemma 1), from which the claim follows. For part (b), differentiation of the leakage rate L_i shows that it is increasing in λ_j if B_i''/B_j'' and B_j''/C_j'' are both constant, as claimed.

Proof of Proposition 8. For part (a), in an interior equilibrium (which implies that $\theta_k > 0$ for $k = i, j$), i's strategic choice of preference $\widehat{\lambda}_i(\lambda_j)$ is determined by its first-order condition $dS_i^*/d\lambda_i = 0$ at Date 1. Using the formula from Lemma 4, this condition can be

written as

$$\widehat{\lambda}_i\theta_i = \theta_i - (1-\lambda_j\theta_i\theta_j)\frac{B_i'}{B_j'}L_i.$$

Differentiating, and using the assumption that B_i'/B_j' is constant, shows that the slope of i's best response curve satisfies

$$\frac{d\widehat{\lambda}_i}{d\lambda_j}\theta_i = \frac{B_i'}{B_j'}\left[\theta_i\theta_j L_i - (1-\lambda_j\theta_i\theta_j)\frac{dL_i}{d\lambda_j}\right].$$

At an interior equilibrium with $(\lambda_i^*, \lambda_j^*) \in (0,1)^2$, this expression can be rearranged as

$$\frac{d\lambda_i^*}{d\lambda_j}\frac{\theta_i}{L_i} = \frac{B_i'}{B_j'}\left[\theta_i\theta_j - \frac{(1-\lambda_j\theta_i\theta_j)}{\lambda_j}\eta_{ij}\right]$$

where the leakage-commitment elasticity $\eta_{ij} \equiv [(dL_i/L_i)/(d\lambda_j/\lambda_j)]_{\lambda_k=\lambda_k^*}$ is evaluated at equilibrium, and from which the result follows immediately. For part (b), using (4) and Lemma 1, we can write the leakage-commitment elasticity $\eta_i \equiv (dL_i/L_i)/(d\lambda_j/\lambda_j)$ as

$$\eta_{ij} = \frac{\lambda_j\left(1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}\right)}{\left(1+\lambda_j\theta_j\frac{B_i''}{B_j''}\right)}\frac{d}{d\lambda_j}\left[\frac{1+\lambda_j\theta_j\frac{B_i''}{B_j''}}{1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}}\right].$$

Differentiating, using the assumption that B_i''/B_j'' and B_j''/C_j'' are both constant, and then simplifying yields:

$$\begin{aligned}\eta_{ij} &= \frac{\lambda_j\left(1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}\right)}{\left(1+\lambda_j\theta_j\frac{B_i''}{B_j''}\right)}\left[\frac{\theta_j\frac{B_i''}{B_j''}\left[1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}\right]-\theta_j\frac{B_i''}{B_j''}\left[1+\lambda_j\theta_j\frac{B_i''}{B_j''}\right]}{\left[1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}\right]^2}\right] \\ &= \frac{\lambda_j\theta_j\frac{B_i''}{B_j''}}{\left(1+\lambda_j\theta_j\frac{B_i''}{B_j''}\right)}\frac{\left[\frac{C_j''}{-B_j''}\right]}{\left[1+\lambda_j\theta_j\frac{B_i''}{B_j''}+\frac{C_j''}{-B_j''}\right]}\end{aligned}$$

Since $\lambda_j^*\theta_j > 0$ in an interior equilibrium, we have that the equilibrium value of the elasticity $\eta_{ij} \in (0,1)$. Observe that the condition from part (a) on the sign of $d\lambda_i^*/d\lambda_j$ still applies as it was derived from weaker assumptions. Since now $\eta_{ij} \in (0,1)$, it follows by inspection that $d\lambda_i^*/d\lambda_j > 0$ if $\lambda_j^*\theta_i\theta_j \geq \frac{1}{2}$, as claimed. Similarly, since $\eta_{ij} > 0$, it is easy to see that $d\lambda_i^*/d\lambda_j < 0$ if $\lambda_j^*\theta_i\theta_j$ is sufficiently small.

Appendix B: Robustness (*not necessarily for publication*)

Generalization to $n \geq 3$ players

Preliminaries. Consider the same setup as in the benchmark model but with $n \geq 3$ players. Let $N = \{1, 2, ..., n\}$ denote the set of players, and let i be a member of this set. Player i's national welfare $\Pi_i = B_i(X) - C_i(X_i)$, where global contributions $X \equiv \sum_{k \in N} X_k$, while global welfare $W = \sum_{k \in N} \Pi_k$, and its true objective $S_i = (1-\theta_i)\Pi_i + \theta_i W$ and strategic objective $\Omega_i = (1 - \lambda_i)\Pi_i + \lambda_i S_i$, where $\theta_i \in [0, 1]$ and $\lambda_i \in [0, \theta_i^{-1}]$. The interaction between players at Date 2, when each chooses effort to maximize her strategic objective Ω_k is an "aggregative game" (Corchón, 1994): each player's objective depends only on its own effort X_k and the (unweighted) sum of all players' efforts X.

Player i's first-order condition for effort at Date 2 can be written as

$$0 = \frac{\partial \Omega_i}{\partial X_i} = \left[B_i'(X) - C_i'(X_i) + \lambda_i \theta_i \sum\nolimits_{j \in N \setminus \{i\}} B_j'(X) \right] \equiv T_i(X_i, X, \lambda_i). \tag{7}$$

The function $T_i(X_i, X, \lambda_i)$ is strictly decreasing in both X_i and X (since $C_k''(\cdot) > 0$ and $B_k''(\cdot) < 0$, respectively for all k). Moreover, whenever $\theta_i > 0$, the function $T_i(X_i, X, \lambda_i)$ is strictly increasing in the strategic preference λ_i. The model thus satisfies Assumptions 1, 2, and 4 of Corchón (1994). Applying Proposition 4 of Corchón (1994) shows that an increase in λ_i leads to (a) a strict increase in X^*, (b) a strict increase in X_i^*, and (c) a strict decrease in X_j^* for all $j \neq i$.

We can recast these results in terms of our model and terminology as

$$\frac{dX^*}{d\lambda_i} = \frac{dX_i^*}{d\lambda_i} + \sum\nolimits_{j \neq i} R_j'(X_i) \frac{dX_i^*}{d\lambda_i} = \left(1 - \sum\nolimits_{j \neq i} L_{ij}\right) \frac{dX_i^*}{d\lambda_i} = (1 - L_i) \frac{dX_i^*}{d\lambda_i} > 0,$$

where $R_j'(X_i)$ is the slope of j's reaction function with respect to i's effort choice. An increase in i's commitment λ_i leads to an increase in its equilibrium effort X_i^* (as in Lemma 2 of our benchmark model). This induces each of the $n-1$ other players to reduce their efforts, that is, player-specific leakage rates $L_{ij} \equiv [-R_j'(X_i)] > 0$. However, the overall leakage rate $L_i \equiv \sum_{j \neq i} L_{ij} \in (0, 1)$ such that global effort rises (as in Lemma 1).

Results. We first derive a generalized version of Lemma 4 from the benchmark model in several steps. With $n \geq 3$ players, the equilibrium impact of a stronger commitment by i on her true objective can be written as

$$\frac{dS_i^*}{d\lambda_i} = \left(\frac{dS_i^*}{dX_i} - \sum\nolimits_{j \in N \setminus \{i\}} \left[\frac{dS_i^*}{dX_j} L_{ij} \right] \right) \frac{dX_i^*}{d\lambda_i}, \tag{8}$$

and we next derive explicit expressions for its individual components.

Step 1. An expression for dS_i^*/dX_i:

$$\begin{aligned}\frac{dS_i^*}{dX_i} &= \frac{d\Pi_i^*}{dX_i} + \theta_i \sum_{j\in N\setminus\{i\}} \frac{d\Pi_j^*}{dX_i} \\ &= \theta_i(1-\lambda_i) \sum_{j\in N\setminus\{i\}} B_j'\end{aligned}$$

where the second equality uses the first-order condition for i from (7) to obtain $d\Pi_i^*/dX_i = (B_i' - C_i') = -\lambda_i\theta_i \sum_{j\in N\setminus\{i\}} B_j'$ as well as the fact that $d\Pi_j^*/dX_i = B_j'$ for any player $j \neq i$.
Step 2. An expression for dS_i^*/dX_j:

$$\begin{aligned}\frac{dS_i^*}{dX_j} &= \frac{d\Pi_i^*}{dX_j} + \theta_i \sum_{k\in N\setminus\{i\}} \frac{d\Pi_k^*}{dX_j} \\ &= (1-\theta_i)\frac{d\Pi_i^*}{dX_j} + \theta_i \left(\frac{d\Pi_j^*}{dX_j} + \sum_{k\in N\setminus\{j\}} \frac{d\Pi_k^*}{dX_j} \right) \\ &= (1-\theta_i)B_i' + \theta_i \left(-\lambda_j\theta_j \sum_{k\in N\setminus\{j\}} B_k' + \sum_{k\in N\setminus\{j\}} B_k' \right) \\ &= (1-\theta_i)B_i' + \theta_i(1-\lambda_j\theta_j) \sum_{k\in N\setminus\{j\}} B_k'\end{aligned}$$

where the second equality uses $\sum_{k\in N\setminus\{i\}} d\Pi_k^*/dX_j = d\Pi_j^*/dX_j + \sum_{k\in N\setminus\{j\}} d\Pi_k^*/dX_j - d\Pi_i^*/dX_j$, and the third equality uses the first-order condition for j from (7).
Step 3. An expression for $\sum_{j\in N\setminus\{i\}} [(dS_i^*/dX_j)L_{ij}]$:

$$\begin{aligned}\sum_{j\in N\setminus\{i\}} \left[\frac{dS_i^*}{dX_j} L_{ij} \right] &= \sum_{j\in N\setminus\{i\}} \left(\left[(1-\theta_i)B_i' + \theta_i(1-\lambda_j\theta_j) \sum_{k\in N\setminus\{j\}} B_k' \right] L_{ij} \right) \\ &= (1-\theta_i)B_i'L_i + \theta_i \sum_{j\in N\setminus\{i\}} \left[(1-\lambda_j\theta_j)L_{ij} \sum_{k\in N\setminus\{j\}} B_k' \right]\end{aligned}$$

where the first equality uses the expression for dS_i^*/dX_j from Step 2, and the second equality uses the definition $L_i \equiv \sum_{j\neq i} L_{ij}$.
Step 4. Another expression for $dS_i^*/d\lambda_i$, to generalize Lemma 4:

$$\frac{dS_i^*}{d\lambda_i} = \left[\theta_i(1-\lambda_i) \sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}} \left[(1-\lambda_j\theta_j)L_{ij} \sum_{k\in N\setminus\{j\}} B_k' \right] \right] \frac{dX_i^*}{d\lambda_i}, \tag{9}$$

which combines the expressions from Steps 1–3 into (8). We use this expression to verify our main results.

(i) *Welfare impact of a small commitment is ambiguous.* Setting $\{\lambda_k\}_{k=1}^n = 0$ in the

expression from (9) yields

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\{\lambda_k\}_{k=1}^n=0} = \left[\theta_i \sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}} \left[L_{ij} \sum_{k\in N\setminus\{j\}} B_k'\right]\right] \frac{dX_i^*}{d\lambda_i}$$

Algebraic manipulation shows that this expression can also be written as:

$$\begin{aligned}
\left.\frac{dS_i^*}{d\lambda_i}\right|_{\{\lambda_k\}_{k=1}^n=0} &= \left[\theta_i \sum_{j\in N\setminus\{i\}} B_j' - B_i'L_i + \theta_i B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}} \left[L_{ij} \sum_{k\in N\setminus\{j\}} B_k'\right]\right] \frac{dX_i^*}{d\lambda_i} \\
&= \left[\theta_i \sum_{j\in N\setminus\{i\}} B_j' - B_i'L_i + \theta_i B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}} \left[L_{ij} \left(\sum_{k\in N\setminus\{i,j\}} B_k' + B_i'\right)\right]\right] \frac{dX_i^*}{d\lambda_i} \\
&= \left[\theta_i \sum_{j\in N\setminus\{i\}} B_j' - B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}} \left[L_{ij} \left(\sum_{k\in N\setminus\{i\}} B_k' - B_j'\right)\right]\right] \frac{dX_i^*}{d\lambda_i} \\
&= \left[\theta_i \sum_{j\in N\setminus\{i\}} B_j' - B_i'L_i - \theta_i \left(\sum_{j\in N\setminus\{i\}} B_j'(L_i - L_{ij})\right)\right] \frac{dX_i^*}{d\lambda_i}
\end{aligned}$$

Next use the covariance identity and define the average marginal benefit among all players $j \neq i$ as $\overline{B}_{-i}' \equiv \frac{1}{(n-1)} \sum_{j\neq i} B_j'$ to obtain

$$\begin{aligned}
\sum_{j\in N\setminus\{i\}} B_j'(L_i - L_{ij}) &= \frac{1}{(N-1)} \sum\nolimits_{j\neq i} B_j' \sum\nolimits_{j\neq i} (L_i - L_{ij}) - (N-1)\cdot cov(B_j', L_{ij}) \\
&= \overline{B}_{-i}'(N-2)L_i - (N-1)\cdot cov(B_j', L_{ij})
\end{aligned}$$

where the second equality uses the definition $L_i \equiv \sum_{j\neq i} L_{ij}$. Using this in the previous expression yields

$$\begin{aligned}
\left.\frac{dS_i^*}{d\lambda_i}\right|_{\{\lambda_k\}_{k=1}^n=0} &= \left[\theta_i(N-1)\overline{B}_{-i}' - B_i'L_i - \theta_i\left(\overline{B}_{-i}'(N-2)L_i - (N-1)\cdot cov(B_j', L_{ij})\right)\right] \frac{dX_i^*}{d\lambda_i} \\
&= \left[\theta_i\left[1 + (N-2)(1-L_i)\right]\overline{B}_{-i}' - B_i'L_i + \theta_i(N-1)\cdot cov(B_j', L_{ij})\right] \frac{dX_i^*}{d\lambda_i}
\end{aligned}$$

We can now state conditions to sign the overall effect of a small commitment, generalizing Proposition 1 from above. Jointly sufficient for $dS_i^* > 0$ are $\theta_i > L_i$, $\overline{B}_{-i}' \geq B_i'$ and $cov(B_j', L_{ij}) \geq 0$. Conversely, if $B_i'/\overline{B}_{-i}'$ sufficiently large and $cov(B_j', L_{ij}) \leq 0$, then $dS_i^* < 0$.

For the global-welfare impact, set $\theta_i = 1$ in the previous derivations to obtain

$$\left.\frac{dW^*}{d\lambda_i}\right|_{\{\lambda_k\}_{k=1}^n=0} = \left[\left[1 + (N-2)(1-L_i)\right]\overline{B}_{-i}' - B_i'L_i + (N-1)\cdot cov(B_j', L_{ij})\right] \frac{dX_i^*}{d\lambda_i}$$

So jointly sufficient conditions for $dW^* > 0$ are $\overline{B}'_{-i} \geq B'_i$ and $cov(B'_j, L_{ij}) \geq 0$, while $dW^* < 0$ if $B'_i/\overline{B}'_{-i}$ sufficiently large and $cov(B'_j, L_{ij}) \leq 0$, thus also generalizing Proposition 2 from the main text.

(ii) *Full commitment is almost never optimal.* Observe first that, since $\lambda_j \theta_j \leq 1$ for any $j \neq i$ the formula from (9) is bounded above according to

$$\frac{dS_i^*}{d\lambda_i} \leq \left[\theta_i(1-\lambda_i) \sum_{j \in N \setminus \{i\}} B'_j - (1-\theta_i) B'_i L_i \right] \frac{dX_i^*}{d\lambda_i}.$$

It follows that $\lambda_i > 1$ cannot be optimal, so we can again henceforth restrict attention to $\lambda_k \in [0,1]$ for all $k = 1, 2, ..., n$. Setting $\theta_1 = \theta_2 = ... = \theta_N = 1$ (for short, $\theta_k = 1 \ \forall k$) in the expression from (9) shows that

$$\left. \frac{dS_i^*}{d\lambda_i} \right|_{\theta_k = 1 \forall k} = \frac{dW^*}{d\lambda_i} = \left[(1-\lambda_i) \sum_{j \in N \setminus \{i\}} B'_j - \sum_{j \in N \setminus \{i\}} \left[(1-\lambda_j) L_{ij} \sum_{k \in N \setminus \{j\}} B'_k \right] \right] \frac{dX_i^*}{d\lambda_i}.$$

So if each player $j \neq i$ is playing $\lambda_j = 1$, then $(dW^*/d\lambda_i)|_{\lambda_j = 1, \forall j \neq i} \geq 0$ for all $\lambda_i \in [0,1]$, and so optimal commitments $\lambda_1^* = \lambda_2^* = ... = \lambda_N^* = 1$ achieve first-best effort levels.

By contrast, setting $\lambda_i = 1$ in (9) gives

$$\left. \frac{dS_i^*}{d\lambda_i} \right|_{\lambda_i = 1} = - \left[(1-\theta_i) B'_i L_i + \theta_i \sum_{j \in N \setminus \{i\}} \left[(1 - \lambda_j \theta_j) L_{ij} \sum_{k \in N \setminus \{j\}} B'_k \right] \right] \frac{dX_i^*}{d\lambda_i}$$

We distinguish between two cases. First, let $\theta_i < 1$. Using our assumption $\lambda_k \leq \theta_k^{-1}$ for all k we have

$$\left. \frac{dS_i^*}{d\lambda_i} \right|_{\lambda_i = 1} \leq -(1-\theta_i) B'_i L_i \frac{dX_i^*}{d\lambda_i}$$

so $(dS_i^*/d\lambda_i)|_{\lambda_i = 1} < 0$ whenever $\theta_i < 1$, and so the optimal commitment $\lambda_i^* < 1$. Second, let $\theta_i = 1$ but $\theta_j < 1$ for at least one other player $j \neq i$, for which then also $\lambda_j^* \theta_j < 1$ by our previous argument. Then we have that

$$\left. \frac{dS_i^*}{d\lambda_i} \right|_{\theta_i = 1} = \left[(1-\lambda_i) \sum_{j \in N \setminus \{i\}} B'_j - \sum_{j \in N \setminus \{i\}} \left[(1 - \lambda_j \theta_j) L_{ij} \sum_{k \in N \setminus \{j\}} B'_k \right] \right] \frac{dX_i^*}{d\lambda_i}$$

and so

$$\left. \frac{dS_i^*}{d\lambda_i} \right|_{\theta_i = 1, \lambda_i = 1} = - \left[\sum_{j \in N \setminus \{i\}} \left[(1 - \lambda_j \theta_j) L_{ij} \sum_{k \in N \setminus \{j\}} B'_k \right] \right] \frac{dX_i^*}{d\lambda_i} < 0$$

since $\lambda_j \theta_j \leq 1$ for any $j \neq i$, and $\lambda_j \theta_j < 1$ for at least one of them. Therefore the optimal commitment $\lambda_i^* < 1$. In summary, if $\theta_k = 1$ for all k, then $\lambda_k^* = 1$ for all k; however, if $\theta_i < 1$ for at least one player i, then $\lambda_k^* < 1$ for all k, thus generalizing our Proposition 3.

(iii) *Optimal commitments can be zero.* Finally we generalize Proposition 4 from the main text to show that (a) if $\theta_j < 1$ for at least one player j (including i) and $B_i'/\overline{B}_{-i}'$ sufficiently large, then $\lambda_i^* = 0$, and (b) if θ_k sufficiently small for all $k = 1, 2, ..., n$, then optimal commitments $\lambda_k^* = 0$ for all k.

For part (a), rewrite the expression from (9) as

$$\begin{aligned}
\frac{dS_i^*}{d\lambda_i} &= \left[\theta_i(1-\lambda_i)\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}}\left[(1-\lambda_j\theta_j)L_{ij}\sum_{k\in N\setminus\{j\}} B_k'\right]\right]\frac{dX_i^*}{d\lambda_i} \\
&= \left[\theta_i(1-\lambda_i)\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i \sum_{j\in N\setminus\{i\}}\left[(1-\lambda_j\theta_j)L_{ij}\left[\sum_{k\in N\setminus\{i,j\}} B_k' + B_i'\right]\right]\right]\frac{dX_i^*}{d\lambda_i} \\
&= \left[\begin{array}{c}\theta_i(1-\lambda_i)\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i B_i' \sum_{j\in N\setminus\{i\}}[(1-\lambda_j\theta_j)L_{ij}] \\ - \sum_{j\in N\setminus\{i\}}\left((1-\lambda_j\theta_j)L_{ij}\sum_{k\in N\setminus\{i,j\}} B_k'\right)\end{array}\right]\frac{dX_i^*}{d\lambda_i}
\end{aligned}$$

By assumption, $\theta_j < 1$ for at least one player j (including i), so we have that $\lambda_k < 1$ for all k by Proposition 3(b) and so also $(1-\lambda_j\theta_j) > 0$ for any $j \neq i$. It follows that the previous expression is bounded above according to

$$\begin{aligned}
\frac{dS_i^*}{d\lambda_i} &< \left[\theta_i(1-\lambda_i)\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i - \theta_i B_i' \sum_{j\in N\setminus\{i\}}[(1-\lambda_j\theta_j)L_{ij}]\right] \\
&= \left[\theta_i(1-\lambda_i)(N-1)\overline{B}_{-i}' - B_i'\left((1-\theta_i)L_i + \theta_i\sum_{j\in N\setminus\{i\}}[(1-\lambda_j\theta_j)L_{ij}]\right)\right]\frac{dX_i^*}{d\lambda_i}
\end{aligned}$$

where second line uses the definition of $\overline{B}_{-i}'$ and does some rearranging. It follows that $dS_i^*/d\lambda_i < 0$ for any λ_i if $B_i'/\overline{B}_{-i}'$ is sufficiently large, such that the optimal commitment $\lambda_i^* = 0$ as claimed.

For part (b), observe that the expression for $dS_i^*/d\lambda_i$ from (9) is bounded above according to

$$\begin{aligned}
\frac{dS_i^*}{d\lambda_i} &\leq \left[\theta_i(1-\lambda_i)\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i\right]\frac{dX_i^*}{d\lambda_i} \\
&\leq \left[\theta_i\sum_{j\in N\setminus\{i\}} B_j' - (1-\theta_i)B_i'L_i\right]\frac{dX_i^*}{d\lambda_i},
\end{aligned}$$

where the first inequality uses our assumption, $\lambda_j\theta_j \leq 1$ for all $j \neq i$, and the second inequality our assumption that $\lambda_i \geq 0$. It follows that, if θ_i is sufficiently small, then $dS_i^*/d\lambda_i < 0$ for all $\lambda_k \in [0, \theta_k^{-1}]^n$, and so i's optimal commitment $\lambda_i^* = 0$. So also, if θ_k is sufficiently small for all k, then optimal commitments $\lambda_k^* = 0$, as claimed.

Cross-country cost spillovers

Preliminaries. Let country k's cost function $C_k(X_i, X_j)$ depend on both countries' effort levels. Country k's national welfare $\Pi_k = B_k(X_i + X_j) - C_k(X_i, X_j)$, and global welfare W as well as its true objective S_k and strategic objective Ω_k are defined as in the benchmark model above ($k = i, j$). The conditions $B_i'' < \partial^2 C_i/\partial X_i \partial X_j < 0$ and $\partial C_i/\partial X_j < 0 \leq \partial^2 C_i/\partial X_j^2$ are sufficient for the model to remain well-behaved, and leakage rates to be strictly positive (but less than 100%).

Country i's first-order condition for its effort choice at Date 2 is given by

$$\frac{\partial \Omega_i}{\partial X_i} = \left(B_i' - \frac{\partial C_i}{\partial X_i}\right) + \lambda_i \theta_i \left(B_j' - \frac{\partial C_j}{\partial X_i}\right) = 0. \tag{10}$$

The leakage rates associated with increased effort by country j is thus

$$L_j \equiv [-R_i'(X_j)] = \left[\frac{-(B_i'' - \partial^2 C_i/\partial X_i \partial X_j) - \lambda_i \theta_i (B_j'' - \partial^2 C_j/\partial X_j \partial X_i)}{-(B_i'' - \partial^2 C_i/\partial X_i^2) - \lambda_i \theta_i (B_j'' - \partial^2 C_j/\partial X_i^2)}\right] \in (0, 1),$$

corresponding to Lemma 1. Using the same arguments as in the proof of Lemma 2, $\text{sign}(dX_i^*/d\lambda_i) = \text{sign}(\partial X_i^*/\partial \lambda_i)$, where

$$\frac{\partial X_i^*}{\partial \lambda_i} = \frac{\theta_i (B_j' - \partial C_j/\partial X_i)}{\left[-(B_i'' - \partial^2 C_i/\partial X_i^2) - \lambda_i \theta_i (B_j'' - \partial^2 C_j/\partial X_i^2)\right]},$$

so that $dX_i^*/d\lambda_i > 0$ whenever $\theta_i > 0$ (since $\partial C_i/\partial X_j < 0$).

Results. In general, the equilibrium impact of a stronger commitment by country i on its true objective can be written as

$$\frac{dS_i^*}{d\lambda_i} = \left[\left(\frac{d\Pi_i^*}{dX_i} + \theta_i \frac{d\Pi_j^*}{dX_i}\right) - \left(\frac{d\Pi_i^*}{dX_j} + \theta_i \frac{d\Pi_j^*}{dX_j}\right) L_i\right] \frac{dX_i^*}{d\lambda_i}$$

Note $d\Pi_i^*/dX_i = (B_i' - \partial C_i/\partial X_i)$, $d\Pi_j^*/dX_i = (B_j' - \partial C_j/\partial X_i)$, $d\Pi_i^*/dX_j = (B_i' - \partial C_i/\partial X_j)$, $d\Pi_j^*/dX_j = (B_j' - \partial C_j/\partial X_j)$, and use the two first-order conditions from (10) to obtain

$$\frac{dS_i^*}{d\lambda_i} = \left[\theta_i (1 - \lambda_i) \left(B_j' - \frac{\partial C_j}{\partial X_i}\right) - (1 - \lambda_j \theta_i \theta_j) \left(B_i' - \frac{\partial C_i}{\partial X_j}\right) L_i\right] \frac{dX_i^*}{d\lambda_i}, \tag{11}$$

which is a generalization of Lemma 4 from the benchmark model.

(i) *Welfare impact of a small commitment is ambiguous.* Setting $\lambda_i = \lambda_j = 0$ in the expression from (11), or using the envelope theorem, shows that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_i = \lambda_j = 0} = \left[\left(\theta_i \left(B_j' - \frac{\partial C_j}{\partial X_i}\right) - \left(B_i' - \frac{\partial C_i}{\partial X_j}\right) L_i\right) \frac{dX_i^*}{d\lambda_i}\right]_{\lambda_i = \lambda_j = 0}.$$

So, generalizing Proposition 1, $dS_i^* > 0$ whenever $(B_i' - \partial C_i/\partial X_j) \leq (B_j' - \partial C_j/\partial X_i)$ and

$\theta_i > L_i$, while $dS_i^* > 0$ if the ratio $(B_i' - \partial C_i/\partial X_j)/(B_j' - \partial C_j/\partial X_i)$ sufficiently large, or if θ_i sufficiently small. By setting $\theta_i = 1$, the impact on global welfare $dW^* \gtrless 0$ according as $(B_j' - \partial C_j/\partial X_i) \gtrless (B_i' - \partial C_i/\partial X_j)L_i$, thus also generalizing Proposition 2.

(ii) *Full commitment is almost never optimal.* Observe first that, since $\lambda_j\theta_j \leq 1$, by assumption, the formula from (11) is bounded above according to

$$\frac{dS_i^*}{d\lambda_i} \leq \left[\theta_i(1-\lambda_i)\left(B_j' - \frac{\partial C_j}{\partial X_i}\right) - (1-\theta_i)\left(B_i' - \frac{\partial C_i}{\partial X_j}\right)L_i\right]\frac{dX_i^*}{d\lambda_i}.$$

It follows that $\lambda_i > 1$ cannot be optimal, so we can again restrict attention to $\lambda_k \in [0,1]$ for $k = i, j$. Setting $\theta_i = \theta_j = 1$ in the expression from (11) shows that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\theta_i=\theta_j=1} = \frac{dW^*}{d\lambda_i} = \left[(1-\lambda_i)\left(B_j' - \frac{\partial C_j}{\partial X_i}\right) - (1-\lambda_j)\left(B_i' - \frac{\partial C_i}{\partial X_j}\right)L_i\right]_{\theta_i=\theta_j=1}.$$

So if country j is playing $\lambda_j = 1$, then $(dW^*/d\lambda_i)|_{\lambda_j=1} \geq 0$ for all $\lambda_i \in [0,1]$, and so optimal commitments $\lambda_i^* = \lambda_j^* = 1$, as in Proposition 3(a). By contrast, setting $\lambda_i = 1$ in (11) shows that $(dS_i^*/d\lambda_i)|_{\lambda_i=1} < 0$ whenever $\theta_i < 1$ or $\theta_j < 1$, and so optimal commitments $\lambda_i^* < 1$ and $\lambda_j^* < 1$, as in Proposition 3(b).

(iii) *Optimal commitments can be zero.* Inspection of (11) shows that $dS_i^*/d\lambda_i < 0$ for all $(\lambda_i, \lambda_j) \in [0,1]^2$ if (a) $\theta_i < 1$ or $\theta_j < 1$ and $(B_i' - \partial C_i/\partial X_j)/(B_j' - \partial C_j/\partial X_i)$ sufficiently large, so that the optimal commitment $\lambda_i^* = 0$, or (b) θ_i and θ_j are both sufficiently small, so that optimal commitments $\lambda_i^* = \lambda_j^* = 0$. These results generalize parts (a) and (b), respectively of Proposition 4.

Other altruistic objective functions

Preliminaries. As explained in the main text, write i's true objective as $S_i = \Pi_i + \widetilde{\theta}_i\Phi_i$, and its strategic objective as $\Omega_i = (1-\widetilde{\lambda}_i)\Pi_i + \widetilde{\lambda}_i S_i$, where $\widetilde{\theta}_i \in [0,1]$ and also let $\widetilde{\lambda}_i \in [0,1]$. To ensure the model remains well-behaved, we assume $\Phi_i(\cdot)$ satisfies $\partial\Phi_i/\partial X_i > 0$ and $\partial\Phi_i/\partial X_j \geq 0$ as well as $[(B_i'' - C_i'') + \partial^2\Phi_i/\partial X_i^2] < 0$ and $\partial^2\Phi_i/\partial X_i\partial X_j \leq 0$.

Player i's first-order condition for her effort choice at Date 2 is given by

$$\frac{\partial\Omega_i}{\partial X_i} = \left(B_i' - C_i'\right) + \widetilde{\lambda}_i\widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_i} = 0. \tag{12}$$

The leakage rates associated with increased effort by j is thus

$$L_j \equiv [-R_i'(X_j)] = \left[\frac{-B_i'' - \widetilde{\lambda}_i\widetilde{\theta}_i(\partial^2\Phi_i/\partial X_i\partial X_j)}{-(B_i'' - C_i'') - \widetilde{\lambda}_i\widetilde{\theta}_i(\partial^2\Phi_i/\partial X_i^2)}\right] \in (0,1),$$

corresponding to Lemma 1. Using the same arguments as in the proof of Lemma 2,

$\text{sign}(dX_i^*/d\widetilde{\lambda}_i) = \text{sign}(\partial X_i^*/\partial\widetilde{\lambda}_i)$, where

$$\frac{\partial X_i^*}{\partial\widetilde{\lambda}_i} = \frac{\widetilde{\theta}_i(\partial\Phi_i/\partial X_i)}{\left[-(B_i'' - C_i'') - \widetilde{\lambda}_i\widehat{\theta}_i(\partial^2\Phi_i/\partial X_i^2)\right]},$$

so that $dX_i^*/d\widetilde{\lambda}_i > 0$ whenever $\widetilde{\theta}_i > 0$ (since $\partial\Phi_i/\partial X_i > 0$).

Results. In general, the equilibrium impact of a stronger commitment by i on her true objective can be written as

$$\frac{dS_i^*}{d\widetilde{\lambda}_i} = \left[\left(\frac{d\Pi_i^*}{dX_i} + \widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_i}\right) - \left(\frac{d\Pi_i^*}{dX_j} + \widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_j}\right)L_i\right]\frac{dX_i^*}{d\widetilde{\lambda}_i}$$

Noting that $d\Pi_i^*/dX_i = (B_i' - C_i')$ and $d\Pi_i^*/dX_j = B_i'$, and using the first-order condition from (12) yields

$$\frac{dS_i^*}{d\widetilde{\lambda}_i} = \left[(1-\widetilde{\lambda}_i)\widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_i} - \left(B_i' + \widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_j}\right)L_i\right]\frac{dX_i^*}{d\widetilde{\lambda}_i}, \tag{13}$$

which corresponds to Lemma 4 from the benchmark model.

(i) *Welfare impact of a small commitment is ambiguous.* Setting $\widetilde{\lambda}_i = \widetilde{\lambda}_j = 0$ in the expression from (13) shows that

$$\left.\frac{dS_i^*}{d\widetilde{\lambda}_i}\right|_{\widetilde{\lambda}_i=\widetilde{\lambda}_j=0} = \left[\left(\widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_i} - \left(B_i' + \widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_j}\right)L_i\right)\frac{dX_i^*}{d\widetilde{\lambda}_i}\right]_{\widetilde{\lambda}_i=\widetilde{\lambda}_j=0}.$$

So $dS_i^* \gtrless 0$ according as $\widetilde{\theta}_i(\partial\Phi_i/\partial X_i) \gtrless [B_i' + \widetilde{\theta}_i(\partial\Phi_i/\partial X_j)]L_i$, with the ambiguous impact corresponding to the result of Proposition 1 (even if $\widetilde{\theta}_i = 1$). It is not difficult to check that the conditions on global welfare $dW^* \gtrless 0$ from Proposition 2 remain exactly the same (given that the commitment is small).

(ii) *Full commitment is never optimal.* Setting $\widetilde{\lambda}_i = 1$ in the expression from (13) shows that

$$\left.\frac{dS_i^*}{d\widetilde{\lambda}_i}\right|_{\widetilde{\lambda}_i=1} = -\left[\left(B_i' + \widetilde{\theta}_i\frac{\partial\Phi_i}{\partial X_j}\right)L_i\frac{dX_i^*}{d\widetilde{\lambda}_i}\right]_{\widetilde{\lambda}_i=1} < 0,$$

such that the optimal commitments satisfy $\widetilde{\lambda}_k^* < 1$, for $k = i, j$, thus providing a stronger version of Proposition 3.

(iii) *Optimal commitments can be zero.* Inspection of (13) shows that $dS_i^*/d\widetilde{\lambda}_i \leq 0$ for all $(\widetilde{\lambda}_i, \widetilde{\lambda}_j) \in [0,1]^2$ if $\widetilde{\theta}_i$ and $\Phi_i(\cdot)$ are such that $\widetilde{\theta}_i(\partial\Phi_i/\partial X_i) \leq [B_i' + \widetilde{\theta}_i(\partial\Phi_i/\partial X_j)L_i]$. Note that a sufficient condition is $(\partial\Phi_i/\partial X_i) \leq B_i'$.