Fair Trade Agreements*

Francesco Passarelli† and Robert W. Staiger‡

May 2024

PRELIMINARY

Abstract

The legitimacy of the world trading system is under growing attack, as challenges to its conformity with norms of fairness and social justice are increasingly voiced by citizens and their governments around the world. Building on concepts of fairness from the philosophy literature and on the economics literature that evaluates the purpose and design of trade agreements, we consider the purpose and design of trade agreements when governments are responsive to the fairness concerns of their constituents. Taking a novel "bottom up" approach to concerns for fairness, we show how these concerns can be formalized in a general and tractable way, and we identify the features of concerns for fairness that would have implications for the purpose and design of a trade agreement. Our findings suggest that as currently designed, the GATT/WTO is well-equipped to allow its member governments to address many, but not all, of the possible trade-related fairness concerns of their citizens. More generally, our findings point to a detailed understanding of real-world perceptions of fairness in trade policy as the key input into the appropriate design of fair trade agreements.

*Brian Moore provided excellent research assistance.

†University of Turin, CESIfo, Baffi Center Bocconi, and Collegio Carlo Alberto.

‡Department of Economics, Dartmouth College; and NBER.

1 Introduction

The legitimacy of the world trading system is under growing attack, as challenges to its conformity with norms of fairness and social justice are increasingly voiced by citizens and their governments around the world. Moreover, these challenges find support in the writings of philosophers and moral theorists, international relations specialists and political scientists. And while economists have argued that many of the design features of the World Trade Organization (WTO), which serves as the constitutional foundation of the world trading system today, can be defended on economic grounds, this defense rarely takes into account the issues of fairness and social justice that animate these challenges. Building on concepts of fairness from the philosophy literature and on the economics literature that evaluates the purpose and design of trade agreements, in this paper we attempt to fill this lacuna by considering the purpose and design of trade agreements when governments are responsive to the fairness concerns of their constituents. We refer to such trade agreements as *fair* trade agreements.

Relative to much of the existing literature on fairness and trade, where concepts of fairness that follow from a set of philosophical and moral axioms are applied to evaluate the fairness of trading arrangements in a "top down" manner, we take a novel "bottom up" approach to the analysis of fair trade agreements that is built on four pillars. First, we treat each citizen of a country as his own moral philosopher, who distinguishes for himself those circumstances when he feels he has been treated fairly from those circumstances when he feels he is being treated unfairly: in effect we view each citizen's opinion of what constitutes fair treatment as his own sovereign right to determine. We therefore take perceptions of unfair treatment as a component of each citizen's preferences and hence as primitives of the model. Second, we assume that when citizens of a country feel that the trade their country is engaged in has been unfair to them, these citizens suffer *psychological harm* ("aggrievement") from this trade that impacts their overall well-being beyond the *material* well-being that they experience in the presence of this trade. Third, we allow that when making their policy choices, governments may be concerned about the perceived unfairness of the trade pressures faced by their citizens – it is in this way that citizens' perceptions about the fairness of trading arrangements may find their way into government preferences over policy. And finally, consistent with the WTO's mandate as a "member-driven" organization, we evaluate the efficacy of a trade agreement's design by its ability to bring its member governments to the efficiency frontier as judged by the preferences of the member governments.

There are several reasons why a government might be concerned about its citizens' perceptions of fairness. First, if the government seeks to maximize the overall utility of its citizens (or particular groups of its citizens) inclusive of their psychological well-being, not just their material well-being, then the government may seek to avoid trade policies that its citizens perceive to be unfair in order to reduce the psychological harm that its citizens suffer. And second, as Passarelli and Tabellini (2017) emphasize, the aggrievement suffered by citizens who feel that they have been treated unfairly can lead to social

conflict, something that a government may find costly for a variety of reasons and which it may wish to avoid.[1] In this case, a government may seek to avoid trade policies that its citizens perceive to be unfair in order to avoid social conflict, even if the government has no direct concern for the psychological well-being of its citizens. We remain agnostic as to the relevance of these various reasons, and simply allow that some (or all) of them may be operative.

From this backdrop, we ask: How do government concerns for fairness impact the purpose of a trade agreement, that is, the "problem" that the trade agreement must "solve" in order to deliver mutual benefits to its member governments? And how do these concerns impact the implied design features of a trade agreement that can serve this purpose? We demonstrate that the answers to these questions hinge on *how* a country's citizens evaluate whether they have been treated fairly. Specifically, we show that what is most critical for the purpose and design of a fair trade agreement is whether the trade pressures that a citizen faces are judged to be fair or unfair on the basis of the trade pressures alone, or rather on the basis of those trade pressures in light of the particular foreign government policies that give rise to them. We also show that the purpose of a trade agreement changes when assessments of fairness are colored by moral stances over the proper use of tariffs by one's own government.

When the fairness of trade pressures is judged without regard to the particular foreign policies that give rise to those pressures – as would be the case for example if citizens of one country feel that it would be unfair for a trading partner to capture a disproportionate share of the gains from trade, or that it would be unfair if trade pressures caused these citizens to experience a lower level of real income than their parents enjoyed – and when moral concerns over own-country tariffs are absent, we find that concerns for fairness are *irrelevant* for the purpose and design of trade agreements, and that the results of Bagwell and Staiger (1999, 2001) extend to this environment without qualification. An implication of these findings is that the basic design of the GATT/WTO is well-equipped to serve its member governments in the presence of such fairness concerns, as Bagwell and Staiger and others have argued is the case in the absence of fairness concerns (see Staiger, 2022, for a review of this literature).

Intuitively, when the fairness of trade pressures is judged on the basis of those trade pressures alone, the international policy externalities that a trade agreement must address in order to deliver mutual gains for its member governments – including the international externalities that are relevant for considerations of fairness – travel through trade flows, and therefore continue to take the form of

[1] For example, perceived unfairness might easily be harnessed by the narrative of populist leaders, leading to political instability and perverse voting behavior (see, for example, Guriev and Papaioannou, 2022, for a survey). Or it might be associated with feelings of ingroup-outgroup identification, leading to costly social conflict (see Iinglehart and Norris, 2016, and Rodrik, 2021). Or it might lead citizens to engage in disruptive forms of political pressure such as protests (see Battaglini, 2017 or Passarelli and Tabellini, 2017). Relatedly, widespread perceptions of unfairness in a society could lead to productivity-reducing "shading" behavior of the kind emphasized by Hart and Moore (2008) and supported by the experimental evidence in Fehr, Hart and Zehnder (2011) when agents feel they have been treated unfairly in the context of private contract negotiations.

a terms-of-trade externality just as Bagwell and Staiger (1999, 2001) showed is the case when fairness concerns are absent. We show that this feature by itself is enough to allow governments to achieve the international efficiency frontier with a GATT-like *shallow approach to integration*, wherein negotiations over tariffs imply market access commitments that are protected from erosion with various GATT/WTO Articles governing permissible non-tariff policy interventions that are otherwise left to the unilateral discretion of each government.

But we show under these conditions that the *purpose* of a trade agreement will nevertheless be impacted even while shallow integration remains viable, if assessments of fairness are colored by moral stances over the proper use of tariffs by one's own government, as when the citizens of some country subscribe to a free-trade ideology when assessing the fairness of their government's trade policy stance. And when this is the case, we show that it may be necessary for a trade agreement to *reduce* trade volumes relative to noncooperative Nash levels rather than increase them in order to generate mutual benefits for the governments. This follows because, while the international externality continues to take the form of a terms-of-trade externality in the presence of these moral concerns, the meaning of this externality is no longer limited to the international cost-shifting interpretation emphasized by Bagwell and Staiger (1999, 2001); and when the problem for a trade agreement to solve involves more than the suppression of international cost-shifting incentives, the inevitability of trade-volume expanding commitments negotiated by governments in a mutually beneficial trade agreement no longer holds. We argue that the presence of these moral considerations could interfere with the efficacy of tariff "bindings," the basic legal commitment to market access in GATT/WTO practice, though we argue that it could also help account for negotiated restrictions on export subsidies that are otherwise hard to explain. And we show that when the fairness of trade pressures is judged on the basis of those trade pressures alone, it is only when these moral considerations are *also* absent from fairness assessments that the purpose and design of a trade agreement is unaffected by fairness considerations and the full results of Bagwell and Staiger (1999, 2001) apply without qualification.

Finally, we demonstrate that it is when the fairness of trade pressures is judged in light of the particular foreign policies that give rise to those trade pressures, as when citizens in some country find per se morally reprehensible a foreign policy allowing production under dangerous working conditions, that concerns for fairness alter the purpose and design of trade agreements in the most fundamental ways. In this case, the viability of shallow integration itself as a means to achieve the international efficiency frontier is disrupted, implying in turn that the GATT/WTO in its present form may be poorly designed to address such fairness concerns. Intuitively, the adoption by one country of policies that citizens in another country consider to be per se morally reprehensible creates a non-pecuniary international externality that is distinct from the terms-of-trade externality that trade flows create and that would otherwise dictate the purpose of a trade agreement. Addressing these fairness concerns then necessitates international negotiations that focus directly on behind-the-border measures, a deep form of integration that would otherwise not be required.

We also investigate fairness concerns as they relate to the maintenance of "personal dignity" in the presence of trade pressures. As Kuziemko, Longuet-Marx and Naidu (2023) observe, such concerns can give rise to a preference for policies that effect "predistribution" – policies that support certain market outcomes in the presence of trade pressures – over policies that effect redistribution via transfers (see also Winkelmann and Winkelmann, 1998). We show that these fairness concerns can impact a government's use of tariffs in both the noncooperative Nash equilibrium and in an internationally efficient trade agreement. But as we demonstrate, these particular fairness concerns have no impact on the purpose or design of a trade agreement, as they neither give rise to a new form of non-pecuniary international externality nor shape the terms-of-trade externality that the trade agreement seeks to address; and therefore they do not warrant special attention when designing a fair trade agreement.

Taken together our findings point to a detailed understanding of real-world perceptions of fairness in trade policy as the key input into the appropriate design of fair trade agreements. And they point to new survey questions that heretofore have not been asked as comprising critical inputs into the best way that trade agreements can be designed to address fairness concerns.

We develop these findings within a two-country two-good general equilibrium neoclassical trade model. To a government objective function that is defined over the material utility of its citizens and where governments can choose trade taxes as in Bagwell and Staiger (1999) and also standards as in Bagwell and Staiger (2001), we append an aggrievement function of the kind described by Passarelli and Tabellini (2017) to capture a government's concerns for its citizens' perceptions of being treated unfairly by trade. A major focus of our paper is the form that this aggrievement function takes. Postponing until a later section the possibility that a citizen's fairness concerns extend to the position of other fellow citizens or other individuals beyond the border, we start with a self-centered notion of fairness. In particular, we assume that a citizen is aggrieved when he feels that he has been treated unfairly by trade, and building on Risse (2007) we assume that feelings of unfair treatment arise whenever one feels that he has not received what he is *owed*. We offer a first formalization of this concept, and in the spirit of Bagwell and Staiger's analysis of government objective functions based on material utility, we then focus on whether the aggrievement function so defined can be written purely as a function of local and world prices or must also include policies directly.

We begin by abstracting from issues of fairness that might arise due to trade's impact on a country's income distribution, and adopt a representative agent model to consider the possibility that *all* of the citizens of a country might feel that they are being treated unfairly in their trade with the other country. The representative agent model is sufficient for us to derive the basic results highlighted above: whether or not fairness concerns have implications for the purpose and design of a trade agreement hinges on whether trade pressures are judged to be fair or unfair on the basis of the trade pressures alone or rather on the basis of those trade pressures in light of the particular foreign government policies that give rise to them, and on whether assessments of fairness are colored

by moral stances over the proper use of tariffs by one's own government.

We then extend our framework to allow for heterogeneous agents within each country and the possibility that the losers from trade competition within a country might themselves feel aggrieved by the trade pressures that they face. In addition to confirming that our results from the representative-agent model generalize to the heterogeneous-agent setting, the heterogeneous agent model allows us to derive new results. Among them is the finding that trade agreements can increase feelings of aggrievement among the citizens of the member countries even when governments are fully responsive to the fairness concerns of their constituents, that is, even when governments negotiate *fair* trade agreements. Intuitively, this is because even for a fair trade agreement, at least part of the purpose – and under the conditions outlined above, the only purpose – of the agreement is to eliminate the international cost-shifting that occurs with policy-induced terms-of-trade improvements and that underpins the Nash policy choices of the member governments. Achieving this purpose inevitably leads to a reduction in trade impediments and an increase in trade volumes; and this may lead to further aggrievement for those citizens that perceived that the level of trade volume they faced was unfair even in the Nash policy equilibrium.

Our heterogeneous agent model also enables us to investigate the fairness concerns that arise when some citizens have a preference for predistribution over redistribution, and we formally confirm the points noted above. In particular, we show that, while these fairness concerns do tilt a government's policy intervention away from transfers and toward the use of tariffs in both the noncooperative Nash equilibrium and in an internationally efficient trade agreement, they have no impact on the purpose or design of a trade agreement and hence do not give rise to the need for special attention in a fair trade agreement.

Finally, we consider the possibility that a citizen's fairness concerns extend to the position of others, either fellow citizens or others beyond the border. We show that the purpose and design of a trade agreement can be impacted by such altruistic feelings, but only when those feelings extend across the border.

Our paper contributes to four literatures. First, we relate to the literature on trade and fairness (e.g., Abbott, 1996, Bhagwati, 1996, Cass and Boltuck, 1996, James, 2005, Davidson, Matusz and Nelson, 2006, Narlikar, 2006, Risse, 2007, Brown and Stern, 2007, Kapstein, 2008, Kurjanska and Risse, 2008, and Risse and Wollner, 2019). As we noted, this literature adopts a top-down approach, where specific concepts of fairness are either assumed or derived from a set of philosophical and moral axioms, and where typically the fairness of trade and trade agreements is then evaluated using these concepts. By contrast, we take a bottom-up approach, allowing different individuals to have different views of what constitutes fair treatment. In this respect, we are consistent with literature in psychology claiming that fairness norms may be substantially different not only across countries but also across citizens of the same country (e.g. Haidt, 2013). We therefore take each citizen's views of fair treatment as primitives of the model – and we evaluate how a trade agreement should be designed when it is understood by the member governments that their citizens will feel aggrieved when they perceive that they are being treated unfairly by trade. A notable

exception in this literature that is closer in spirit to our paper is the paper by Davidson, Matusz and Nelson (2006), who adopt a positive perspective and consider a small-country median-voter model of trade policy determination to explore the implications of widely held notions of fairness for equilibrium trade policy. They do not, however, consider the implications for the design of trade agreements, which is our focus here.

Second, we relate to Passarelli and Tabellini (2007), who introduce notions of fairness into a model that directly maps government policies into feelings of unfair treatment, and ultimately into political unrest and protests that are costly for a government, and who then study the determination of public policy within this setting. Like them, we do not impose concepts of fairness from the top down. But unlike them, we focus specifically on concerns for fairness as these concerns relate to international trade, and on the implications of these concerns for the design of trade agreements; and to do this we delve deeply into the structure of the mapping from material outcomes and policies into citizens' perceptions of unfair treatment, an understanding of which we show is a critical input into the design of a trade agreement in the presence of fairness concerns.

Third, we relate to Bagwell and Staiger (1999, 2002), who study the purpose and design of trade agreements, and to Bagwell and Staiger (2001) and Staiger and Sykes (2011, 2021), who extend the study of the purpose and design of trade agreements to the treatment of domestic policies and the possibility of shallow integration. Like us, these papers focus on identifying the underlying purpose of a trade agreement and characterizing design features that can achieve that purpose, but these papers do not consider issues of fairness. We establish conditions under which the results from this literature extend without qualification to a world where fairness considerations are important, and conditions when these results must be modified in such a world.

And fourth, our paper relates to the work of a group of international legal scholars who seek to interpret the implications for the design of international trade agreements of the heightened prominence of "non-economic" concerns and the shifting norms governing the authority to tax and regulate international commerce that these concerns imply (e.g. Pauwelyn and Sieger-Gasser, 2024, Shaffer, 2024, Meyer, forthcoming). Like these authors, we seek to assess when addressing these concerns creates new requirements for the design of trade agreements and when it does not, though our focus on fairness is a subset of the concerns that these authors have in mind. Unlike these authors, we develop a formal framework within which to make this assessment.

The rest of the paper proceeds as follows. The next section provides some background on the literature concerned with fairness issues related to international trade relations. Section 3 sets out the trade model within which our analysis is carried out, and introduces an aggrievement function which describes the structure of the mapping from material outcomes and policies into citizens' perceptions of unfair treatment. Sections 4 and 5 contain our core analysis of fair trade agreements, first for the representative agent model and then for the heterogeneous agent model. Section 6 extends our analysis to consider concerns over the fair treatment of others. Section 7 concludes.

2 Fairness and Trade

To set the stage for the analysis to follow, we first provide some background to the literature on fairness. For economists unfamiliar with this literature, a useful caution is provided by Suranovic (2000), who observes:

> The literature on fairness is diverse, multi-disciplinary, and often impenetrable. The concept itself overlaps with many other normative principles such as justice, equity, law and even morality. As such, one cannot simply pick up a book or article and quickly discover what fairness means or how to distinguish between the various normative principles. And yet, at the same time, everyone seems to have an inherent sense of what fairness is. (p 283)

As this quote suggests, the literature on fairness is not easily digested, and it is not our purpose here to provide a comprehensive summary. Rather, without attempting to survey, summarize or synthesize this literature, in this section we simply distill from the literature a few key concepts and definitions that we will use to guide our subsequent analysis. As we will see, even this seemingly simple task turns out to be rather involved.

What is fairness? First, what is fairness?[2] Moral judgements derive from a comparison between a fair "reference transaction" and the actual transaction (e.g. Kahneman, Knetsch, and Thaler, 1986). The former refers to how the transaction should to be in a counterfactual fair world, the latter is how the transaction actually is. A fair transaction then shapes an individual's expectations of fair treatment. Since the fair transaction carries moral weight, it gives rise to what Risse (2007) calls an individual's *stringent claims*, i.e., what the individual is owed (or, as we will apply this concept below, what the individual *feels* he is owed). With reference to the concept of stringent claims, Risse provides a useful statement of what fairness is, and what it is not:

> ... Discussions about fairness often concern distributions of goods (for example, inheritances or kidneys) or burdens (for example, taxes or layoffs), as well as processes governing such distributions. While 'fairness in trade' is more abstract than such scenarios, similar issues arise. One could assess such distributions in many ways: one may ask which one maximizes welfare, inflicts the least maximal harm, or best satisfies external goals. Yet fairness evaluates distributions in a special way.
>
> Fairness ensures people receive *what they are owed*. I refer to demands people have because they are owed something as *stringent claims*. Distributions of burdens or benefits are not fair (or unfair) *merely* because they meet (or violate) any criterion just mentioned. They are unfair only if they fail to deliver what people are owed.

[2] We follow Konow (2003) and use the terms fairness and justice interchangeably here.

> Philanthropists are not unfair if they give more to one university while another has bigger needs (neither having a stringent claim). ... (Risse, 2007, pp 357-358, emphasis in the original)

In what follows we will evaluate the fairness of a trade agreement on the basis of how effectively it delivers to people what they feel they are owed, and thereby fulfills their stringent claims.

Where do stringent claims come from? Second, if fairness is achieved when people's stringent claims are satisfied, where do these stringent claims come from? Here we depart from the literature's normative "top-down" approach to this question, in which moral theories of fairness are proposed and defended on normative grounds and then employed to derive the stringent claims and moral imperatives that are implied by those theories. Instead we take a positive "bottom-up" approach and assume that each citizen acts as his own moral philosopher and defines his own stringent claims, which if not met will lead to feelings of aggrievement. By allowing fairness to be defined in the eye of the beholder and switching the focus of the analysis to how aggrievement can be addressed by a trade agreement, our approach sidesteps the thorny normative question regarding what is the "right" notion of fairness in trade.

Our approach can be contrasted with the more normative approach to the fairness-in-trade question taken in the influential Bhagwati and Hudec (1996) edited volume on fair trade and harmonization. For example, in his introduction to Volume 2, Hudec (1996) observes that

> ... the norms by which current political institutions tend to judge international trade issues leave a great deal to be desired in terms of coherence, consistency, and objectivity. ... The first step in the process of paring down conflicting value claims to their essentials should be a very critical look at all of the fairness norms and other kinds of moral imperatives underlying this conflict. (pp 16-17)

Rather than attempting to advance the normative analysis of fairness in trade as Hudec suggests, we follow Suranovic (2000) and especially Davidson, Matusz and Nelson (2006) in providing a positive analysis of these issues, with an emphasis on their implications for the design of trade agreements.

What forms might stringent claims take? In principle, we allow them to take any form (they are the sovereign right of the citizen to define for himself), but we can make use of the literature on moral philosophy to categorize the possible claims. Broadly speaking, these claims may concern *distributive justice* for the individual, that is, a concern with the fairness of outcomes; and/or they may concern *procedural justice* for the individual, that is, a concern with the fairness of the processes that deliver the outcomes (Konow, 2003, Risse, 2007). More specifically, it is useful to provide a mapping from the major theories in the moral philosophy literature to these two broad categories of stringent claims, with the understanding that we will not impose restrictions on an individual's

stringent claims by taking a stand on which of these moral theories is "valid," but rather make use of this mapping simply to illustrate the sorts of moral philosophizing that would go into determining an individual's stringent claims.

Much of the literature on fairness in trade emphasizes the division of the gains from trade across countries, with the fairness view being justified by principles of structural equity and based on different notions of national sovereignty. These concerns for fairness typically regard the use of power by rich countries to distort distribution in their favor or the exploitation of low wages by poor countries to hurt the rich ones (the so-called Pauper Labor Argument). Discussion centers around whether international trade rules should be set up to ensure structural equity (i.e., a reasonably acceptable distribution of benefits among countries, as in James, 2014), or whether rationally self-interested bargaining leading to contribution-based distribution is fair, with no need for distributional rules (Gauthier, 1986). This literature typically abstracts from the individual's point of view, striving instead for a definition of fairness that can be universally applied to all countries and individuals.

Yet individuals may develop feelings of unfairness for their own position or the position of other individuals, independently of the division of gains across countries (Suranovic, 2000; Miller, 2017). Recent empirical research suggests that people have moral concerns about the distribution of gains not only across countries but also across individuals, and their concerns are not only related to outcomes but also to government policies. Stantcheva (2023) finds that citizens express strong concerns about the adverse distributional consequences of international trade, as well as the policies designed to compensate the losers. Kuziemko et al. (2023) show that people may prefer predistribution (e.g., keeping their own job) over redistribution (e.g., losing their job but receiving compensatory cash transfers) when it comes to trade policies, indicating that they might feel entitled to be protected from trade pressure rather than compensated in case of loss, a conclusion that is also supported by the findings of Winkelmann and Winkelmann (1998). Di Tella and Rodrik (2022) find that individuals demand more import protection after a trade shock when the trading partner is a developing country compared to a developed country. This may be due to the perception that developing countries' policies are more unfair.

So *what* does an individual think he is owed, and *why* does he think he is owed it? On the question of what, there are three possibilities to consider.

First, an individual may think he is not owed anything at all. This is the utilitarian individual in neoclassical models of trade, where agents are assumed to have no stringent claims whatsoever. Any transaction is then fair, except for those that are clearly repugnant (e.g., Roth, 2007). The individual's goal is to maximize his material utility with no moral sentiments about how utility is produced or distributed.

Second, an individual might think he is owed a minimum level of material utility that he deems fair, without regard to how that level is delivered. This is also a utilitarian individual, but one who would become aggrieved and suffer psychological harm if his material utility falls below a certain level. He attaches no moral sentiments to whether his income comes from, say, a $100 transfer or

a new job that provides a net $100 income. In other words, this individual has no preference for predistribution over redistribution. We say that he is guided by distributive justice goals.[3]

Third, an individual might have moral sentiments not only about a certain amount of material utility that he feels he is owed, but also about how the material utility is delivered. He might have moral sentiments about home-government policies such as tariffs, either opposed to or supportive of tariff intervention on moral grounds; he might have moral sentiments about transfers, feeling degraded when he is on the receiving end; or he might have moral sentiments about safety standards in his workplace. He might have moral sentiments about whether he will be forced to change jobs or move to a new community in order to generate his material utility. He might even have moral sentiments about foreign-government policies such as export subsidies or standards beyond the border.[4] This individual is guided by procedural justice goals, implying that he has moral sentiments not only on outcomes (material utility levels) but also on processes (home or foreign policies and/or the workplace attributes that they induce). This is the individual attaching moral-dignity sentiments to the way he is compensated by his government for a loss from trade, or having a taste for a "level playing field" in international competition.

Why does the individual think he is owed something? In answering this question we can align an individual's stringent claims with existing theories of moral justice and provide an underlying interpretation of those claims. Using Konow's (2003) categories of leading justice theories we envisage three possibilities: Equity and Need, Equity and Desert, and Context.[5]

First, the individual's moral sentiments may align with Equity-and-Need justice principles, which incorporate concerns for the least-well-off members of society. These principles may derive from the idea that an individual behind a veil of ignorance would be willing to sign a social contract providing a form of social insurance against negative shocks (e.g., Rawls, 1971, Binmore, 1994, Harsanyi, 1975). Therefore losers from trade might feel entitled to compensatory transfers, while trade winners would deem it fair to contribute to those transfers.

[3] This is also the individual described by Fehr and Schmidt (1999) and Charness and Rabin (2002). He has moral feelings about how income should be distributed and might dislike inequity. Those feelings affect his utility but he has no moral sentiments about the policies to achieve redistribution.

[4] Shaffer (2024) vividly illustrates such moral sentiments with a quote from William Ellery Channing's 1836 *Tribute to the American Abolitionists*:

> "The South says, that slavery is nothing to us at the North. But, *through our trade*, we are brought into constant contact with it; we grow familiar with it; still more, we thrive by it; and the next step is easy, *to consent to the sacrifice of human beings by whom we prosper*." (italics in original) -William Ellery Channing, "Tribute to the American Abolitionists for their Vindication of Freedom of Speech," American Anti-Slavery Society (New York 1861) [1836]

[5] Konow (2003) surveys the literature on both distributive and procedural justice, and considers how accurately the theories of justice in each of these categories describe the actual fairness preferences of people. See also Miller (1992), who surveys the evidence across these categories regarding people's actual social justice beliefs focusing on distributive justice.

Stringent claims with Equity and Need are guided by distributive justice rather than procedural justice. And the distributive justice concerns at issue could be either national (within-country income distribution) or transnational (cross-country division of the gains from trade).

Second, the individual's moral sentiments may align with Equity-and-Desert theories of justice, which are based on proportionality and individual responsibility. Outcomes are fair as long as they result from effort, luck, or choice (Buchanan, 1986). According to these principles of justice, foreign export subsidies, for instance, or weak labor standards beyond the border, may be considered unfair because they distort competition and allow foreign firms to gain without effort, while gains from technological breakthroughs are considered fair because they result from effort and responsibility. Therefore an individual might think he is owed $100 as compensation for a loss he suffered due to foreign policies distorting competition, while he might think he is owed nothing if that loss was due to a technological breakthrough in the foreign country. Claims aligning with Equity and Desert are thus guided by procedural justice goals.

Third, the individual's stringent claims may align with the family of justice theories based on Context. Context theories do not generate a distributive principle but highlight the dependence of justice evaluation on the context, such as historical terms of transactions or the type of good being distributed (e.g., Kahneman, Knetsch, and Thaler, 1986). Various contextual factors can shape an individual's stringent claims. For instance, he might feel that an outcome is fair as long as it reflects past transactions. This could explain why displaced workers might think they are owed their past job or their past level of protection from imports. Context-dependent justice allows for cultural or historical factors to affect claims of fairness, which can vary across trading countries. This would explain why citizens of country A might consider social policies in country B unfair, while citizens in country B would not. Thus, claims aligning with Context are guided primarily by procedural justice goals.[6]

Who owes what to whom? Finally, if citizen i feels he is owed something, who owes it to him? If citizen i resides in the domestic country and the thing he feels he is owed is controlled by the foreign government, does the foreign government owe something directly to domestic citizen i based on moral grounds?

In answer to this question, we will adopt a view of stringent claims that maintains consistency with the sovereignty of nations, and assume that a citizen's stringent claims are always claims on his *own* government, never on the governments of other countries, even if these claims are generated by the policy actions of foreign governments. Hence, for example, it is not that domestic citizen i feels morally that he is owed something from the foreign government when the foreign government adopts a policy stance that citizen i feels treats him unfairly; instead this foreign policy stance causes domestic citizen i to feel on

[6] We say "primarily" because in the cases that Context theories typically have in mind, the fair "transaction" would include the process that led to the transaction and not just the outcome. But it is possible to imagine scenarios in which the transaction concerns only the outcome, in which case distributive justice goals would apply.

moral grounds that he is owed something from his own government to address his stringent claim regarding this foreign policy.

We adopt this view as a positive matter, consistent with the positive emphasis of our analysis. Whether claims on a foreign government are justifiable from a *normative* perspective is a matter of debate in the literature. Risse (2007) and Kurjanska and Risse (2008) provide a particularly clear statement of the various positions in this debate.

Based on what Risse (2007) calls a "Strong Westphalian View," countries are sovereign in determining their own trade policies and their own standards and social costs of production, unless their production involves atrocious activities (such as slavery) or the production process itself causes material harm to the other country (such as cross-border pollution). According to this view, aside from exceptional circumstances countries owe nothing to each other: a foreign government does not owe anything to domestic citizens, nor does the domestic government owe anything to its own citizens based merely on the policy choices of the foreign government. Under the Strong Westphalian View, then, stringent claims on a foreign government are not justifiable as a normative matter.

By contrast, according to Risse (2007) the so-called "Moderate Westphalian View" does justify stringent claims on a foreign country under certain conditions. For example, under this view foreign workers have a stringent claim on a domestic country from a normative perspective if the domestic country trades with the foreign country while the foreign workers are oppressed. At the same time, according to this view domestic citizens have justifiable stringent claims to protection by their government, if they have been harmed by trade pressures associated with foreign trade policies or foreign social standards that are at odds with domestic social standards.

The citizens of one country might also have moral feelings about the income of the poor beyond their borders. They might feel that the foreign government owes its citizens a better standard of living. Such a view about income levels and distribution in the foreign country implies a certain degree of altruism (or egalitarian concern) by the domestic citizen towards foreign citizens. It also implies that, as a normative matter, a citizen in the domestic country may have stringent claims against a foreign-country government. According to Kurjanska and Risse (2008), such a claim would be justifiable as a normative matter only based on a so-called "Weak Westphalian View," which holds that every country is subject to "constraints in fairness that limit how it can determine the social costs of production," and that as a normative matter trade policies should be "devised in such a way as to be consistent with duties to poor countries."

Our positive approach to the question of who owes what to whom allows us to sidestep the normative debate in the fairness literature on this question.[7]

[7] Below, we also consider the related concepts of "production jurisdiction" and "consumption jurisdiction" introduced by Meyer (forthcoming), where the former is consistent with Risse's (2007) notion of Strong Westphalian View, while the latter aligns with the Weak Westphalian View described in Kurjanska and Risse (2008). Our positive approach also allows us to avoid a prominent issue in the literature on standards harmonization across countries (e.g., Bhagwati and Hudec, 1996) which is logically separable from the issue of fairness in trade

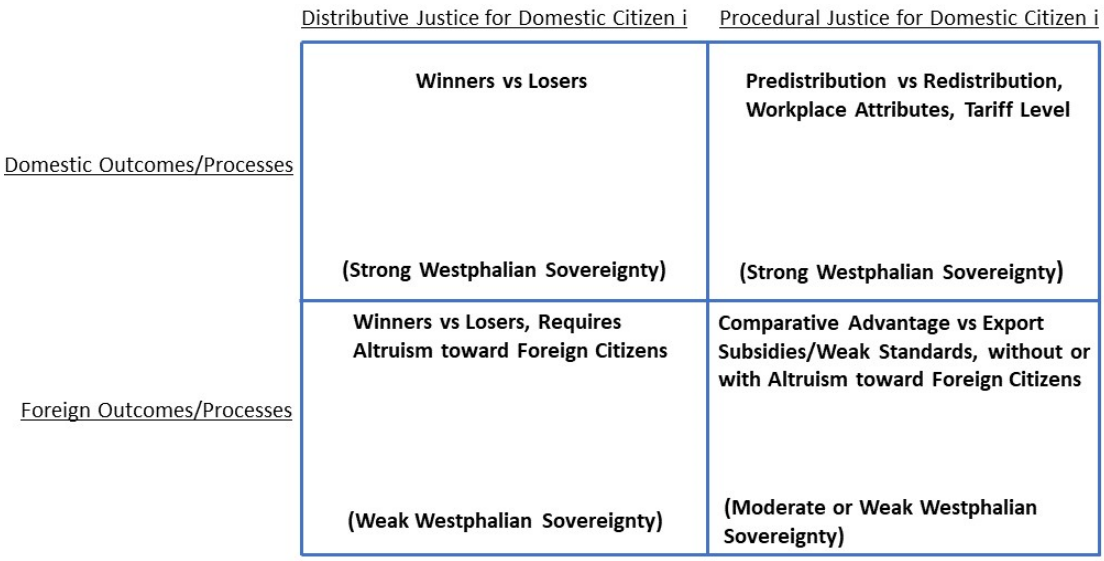

Figure 1: A Taxonomy of Moral Sentiments

A taxonomy of moral sentiments As mentioned, we employ these concepts in a bottom-up way, allowing each citizen to be his own moral philosopher and determine his own stringent claims. And as we noted, the different moral theories we have touched on above may guide the citizen in coming up with his stringent claims (or possibly with his own moral theory).

Figure 1 provides a taxonomy of where a citizen's moral sentiments may land, for illustrative purposes taking the perspective of domestic citizen i. The figure distinguishes between an individual's moral feelings about domestic outcomes/processes (first row) and foreign outcomes/processes (second row). It also indicates whether these claims are independent of policies and context and relate to distributive justice concerns (first column) or dependent on policies and/or context and relate to procedural justice concerns (second column). The meaning of the term in parentheses in each box of Figure 1 is that, if citizen i's

per se: Does country B have any obligation to a citizen of country A? More specifically, can a citizen of country A, based on his own moral evaluations, request that country B improve its safety standards for workers in that country? For example, the countries of the European Union have made significant progress in integrating their policies by voluntarily relinquishing a substantial part of their national sovereignty. This renunciation has made extensive forms of harmonization possible, including the introduction of principles of mutual recognition that effectively limit the sovereignty of member countries, requiring them to accept the policies of others. Harmonizing policies based on moral issues would necessitate a shared vision of moral principles on which to establish common rules. On what moral principle can country A ask country B to align its policies with its own? Our positive approach, based on the assumption that countries maintain full sovereignty as is the case for WTO members, need not address these issues.

stringent claims were about what is in the box, these stringent claims would be consistent with the normative view of fairness listed inside the parentheses in that box as defined by Risse (2007) and Kurjanska and Risse (2008).

As we will describe below, the taxonomy in Figure 1 is suggestive of a set of stylized questions that might be posed to individuals in order to determine the relevant dimensions of their moral sentiments and position them in this taxonomy. We will not in this paper operationalize these questions with a survey that might tell us where in the taxonomy real-world moral sentiments reside. Rather, our goal in this paper is simply to provide a mapping from this taxonomy of moral sentiments to the design features of a trade agreement that could be said to be fair in light of these moral sentiments.

3 The Basic Trade Model

In this section we introduce the basic model of the world economy that will provide the framework for our subsequent analysis. We begin by laying out the notation and basic relationships for the world economy, and then turn to a detailed treatment of the aggrievement functions which captures the psychological components of welfare that we highlight in our treatment of fairness and trade.

3.1 The Model World Economy

We consider a simple general-equilibrium neoclassical trade model with two countries and two goods. Throughout we use an asterisk "*" to denote foreign-country variables. Markets are perfectly competitive. The home country exports good y to the foreign country in exchange for imports of good x from the foreign country.

Let $p \equiv p_x/p_y$ denote the home country's local relative price of good x to good y, and let $p^* \equiv p_x^*/p_y^*$ be the local relative price in the foreign country. The world relative price of good x to good y can then be defined as the ratio of exporter prices, $p^w \equiv p_x^*/p_y$, and p^w gives the terms of trade between the two countries: a rise in p^w corresponds to a worsening of the home country's terms of trade and an improvement in the terms of trade for the foreign country. The home and foreign countries can each impose an ad valorem import tariff, t and t^*, respectively. If t or t^* is negative, then this is an import subsidy, which by Lerner Symmetry can be equivalently thought of as an export subsidy. Let $\tau \equiv 1 + t$ and $\tau^* \equiv 1 + t^*$. Sometimes we will refer to τ and τ^* as the home- and foreign-country's tariffs, directly. By the arbitrage condition that must hold with strictly positive trade volumes (which we assume throughout), the home country's local relative price is then

$$p = \tau p^w \equiv p(\tau, p^w) \tag{1}$$

and the foreign country's local relative price is

$$p^* = p^w/\tau^* \equiv p^*(\tau^*, p^w). \tag{2}$$

We assume that the government of each country redistributes to its citizens in a lump-sum fashion the tariff revenue it collects. More generally we allow for the possibility that a government may have lump-sum instruments to make arbitrary transfer payments across its citizens. We let T^i denote the transfer that domestic citizen i receives beyond his share of tariff revenue (tax if negative), with T^{*i} defined analogously for the foreign country. To maintain focus on the main points, we further assume that the citizens of a country share identical and homothetic preferences over consumption of x and y, so that any redistribution of income among them has no impact on the country's aggregate demands. We also allow each government to set its standard, s for the domestic country and s^* for the foreign country. Following Bagwell and Staiger (2001), these standards are best thought of as production standards that stipulate a minimum working age or a workplace health or safety regulation.[8]

Each country's production possibilities frontier is pinned down by its technologies, its factor endowments, and its standard (which could impact the factor endowments or technologies that can be legally employed within its borders), and production in each country occurs at the point on the country's production possibility frontier at which the (negative of the) local relative price equals the slope of the production possibilities frontier (the marginal rate of transformation between the two goods). And given the preferences of each country's citizens, each country's aggregate demands are pinned down once its local relative price (which determines the country's level and distribution of real factor incomes across its citizens and the tradeoff its citizens face in consumption) and world relative price (which together with the local relative price then determines tariff revenue) are known.

Hence, for any world relative price p^w and any (non-prohibitive) tariffs τ and τ^*, the home and foreign local relative prices are determined; and for given standards s and s^*, technologies, endowments and preferences in the two countries, the home-country import demand for good x and the foreign-country export supply of good x is then also pinned down. The equilibrium relative world price $\tilde{p}^w(s, \tau, s^*, \tau^*)$ is then determined by the market clearing condition that equates home-country imports of good x, M, with foreign-country exports of x, E^*, given by

$$M(s, p(\tau, \tilde{p}^w), \tilde{p}^w) = E^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w), \tag{3}$$

with market clearing for good y then guaranteed by Walras' Law.[9] And under standard conditions to rule out the Lerner and Metzler paradoxes, we have

$$\begin{array}{rcl} \dfrac{\partial \tilde{p}^w(s, \tau, s^*, \tau^*)}{\partial \tau} & < & 0 < \dfrac{\partial \tilde{p}^w(s, \tau, s^*, \tau^*)}{\partial \tau^*} \text{ and} \\ \dfrac{dp(\tau, \tilde{p}^w(s, \tau, s^*, \tau^*))}{d\tau} & > & 0 > \dfrac{dp^*(\tau^*, \tilde{p}^w(s, \tau, s^*, \tau^*))}{d\tau^*}. \end{array} \tag{4}$$

[8] Though see Bagwell and Staiger (2001, note 8) for an interpretation that extends to consumption standards; and see Staiger and Sykes (2011) for further elaboration on the extension to consumption standards.

[9] Our assumption of identical and homothetic preferences within each country implies that within-country transfers are irrelevant for market clearing conditions, and hence the transfers T^i and T^{*i} do not enter into the market-clearing world price function.

For now we do not impose any structure on the signs of the impacts of standards on world prices (e.g., we allow that a weaker standard if applied to its import-competing sector could improve the domestic country's terms of trade while if applied to its export sector this weaker standard could worsen the domestic country's terms of trade, and similarly for the foreign country).

Finally, a word on how we will approach the issue of fairness in our model world economy. We will assume that governments themselves are amoral, and make no judgements of their own as to the fairness of a given situation. Rather, as we discussed in section 2 and formalize below, it is the citizens of each country who have moral sentiments and make judgements about fairness. Governments, however, may be responsive to the moral judgements of their citizens, either because the psychological harm that their citizens experience in a world that these citizens consider unfair is given direct weight by the government objective function, or because the costs of aggrievement associated with unfair treatment in terms of social disruption, protests, shading behavior and the like are given weight in the government objective function. To formalize this, we next turn to a development of our aggrievement function.

3.2 The Aggrievement Function

We now describe the novel psychological aspects of welfare that we introduce into the model. We capture these aspects through an *aggrievement function* that maps policies and material outcomes into citizens' perceptions of unfair treatment and the psychological harm that they suffer as a result. For simplicity, we develop our specification of the aggrievement function from the perspective of the domestic country.

Material utility We begin with domestic citizen i's material utility, which we represent by citizen i's indirect (material) utility function

$$\begin{aligned} v^i &= v(p(\tau,\tilde{p}^w), I^i(s,p(\tau,\tilde{p}^w),\tilde{p}^w)+T^i,s) \\ &\equiv V^i(s,T^i,p(\tau,\tilde{p}^w),\tilde{p}^w(s,\tau,s^*,\tau^*)), \end{aligned} \tag{5}$$

where I^i is citizen i's factor income plus his share of tariff revenue, measured in units of the domestic-country export good y, where T^i is the transfer (transfer if positive, tax if negative) received by citizen i beyond his share of tariff revenue, also measured in units of good y, and where there is no i subscript on the domestic indirect utility function $v(\cdot)$ due to our assumption that preferences over the consumption of x and y are homothetic and identical across the citizens of the domestic country.[10] The top line in (5) expresses citizen i's indirect utility as a function of citizen i's income and the prices he faces, but also includes the

[10] Formally, domestic national factor income plus tariff revenue all measured in units of y can be written as $I(s,p,\tilde{p}^w) \equiv pQ_x(s,p) + Q_y(s,p) + R(s,p,\tilde{p}^w)$ where $R(s,p,\tilde{p}^w) \equiv (p-\tilde{p}^w)M(s,p(\tau,\tilde{p}^w),\tilde{p}^w)$ is tariff revenue collected by the domestic government. $I^i(s,p,\tilde{p}^w)$ is then citizen i's equilibrium share of domestic national factor income plus tariff revenue, which is determined by s, $p(\tau,\tilde{p}^w)$ and $\tilde{p}^w$ and the share of tariff revenue that citizen i receives.

domestic standard s to capture possible health and safety aspects associated with the standard that may impact the material utility of the citizens of the domestic country (e.g., s is a workplace safety standard that impacts citizen i's probability of death while on the job, or s is a pollution standard that impacts the level of local "eye sore" pollutants that detract from citizen i's utility).[11] Notice that in terms of material utility, factor income and government transfers are assumed to be perfect substitutes; it is in the psychological aspects of welfare that we consider next that this perfect substitutability may not hold, and instead a preference for "predistribution" (factor income) over redistribution (government transfers) may arise on the grounds of maintaining one's "personal dignity."[12] The bottom line in (5) rewrites the indirect utility function of citizen i in an equivalent form that emphasizes the role of local and world prices in determining material utility (as in Bagwell and Staiger, 1999, 2001).

By the properties of the indirect utility function, v is strictly increasing in its second argument $I^i(s, p(\tau, \tilde{p}^w), \tilde{p}^w) + T^i$, and we will assume that v is also concave in this argument; by implication, V^i is then increasing and concave in T^i. Moreover, notice that an increase in $\tilde{p}^w$ corresponds to a worsening terms of trade for the domestic country, and with the other arguments of V^i held fixed this simply implies less tariff revenue to be redistributed back to citizen i and his fellow citizens. Similar to Bagwell and Staiger (1999, 2001), we therefore impose additional minimal structure on citizen i's material utility V^i, and assume that it is strictly decreasing in $\tilde{p}^w$:[13]

$$\frac{\partial V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))}{\partial \tilde{p}^w} < 0. \qquad \text{(Assumption 1)}$$

We will impose the analogous structure on the indirect material utility of foreign citizen i.

Stringent claims To develop our aggrievement function, we build on Risse's (2007) concept of *stringent claims*, but we take a different approach from Risse in making use of this concept, an approach that is more amenable to formalization. Ultimately, our goal is to put some structure on the question whether fairness in trade is judged on the basis of the trade pressures alone, or rather on the basis of trade pressures in combination with the specific government policies that give rise to those trade pressures.

In particular, we treat citizen i's stringent claims - what he feels he is owed - as deriving from a *sovereign right* to hold his own view of fair trade, his own

[11] We assume that the chosen domestic standard impacts the material utility of all domestic citizens in the same way, but our results do not depend on this assumption.

[12] For simplicity we are treating citizen i's share of tariff revenue as different from the government transfer T^i that citizen i might also receive in terms of any possible personal dignity effects, but none of our results would change if we included citizen i's share of tariff revenue in T^i as well.

[13] Bagwell and Staiger (1999, 2001) impose this structure directly on government objective functions. For our purposes, it is convenient to impose this structure on the indirect (material) utility function of individual citizens, and then let the government objective functions inherit this structure.

moral sentiments about how trade transactions should be allowed to impact his life, in the same way as it is his sovereign right to have his own material preferences over consumer goods. We focus on a *self-centered* concept of fairness as in Fehr and Schmidt (1999), abstracting from altruistic concepts of fairness that would include concerns over whether others are treated fairly.[14] Finally, we take the equilibrium relative world price $\tilde{p}^w(s, \tau, s^*, \tau^*)$ as a summary of the challenges and opportunities that trade presents to the domestic country, which as a shorthand we refer to as "foreign trade pressures."

Inspired by Risse (2007), we then assume that citizen i will feel that he is being *treated unfairly by trade* – and citizen i will therefore be *aggrieved* – if and only if citizen i fails to receive what he feels he is *owed* and sees foreign trade pressures as a major cause of this state of affairs. Our focus on trade-related issues of fairness to the exclusion of fairness issues that arise in other domains reflects our interest in trade agreements, and the view that even fair trade agreements cannot be expected to address all issues of unfairness in the world. As we mentioned in section 2, we adopt a view of stringent claims which is consistent with the maintenance of national sovereignty and assume that a citizen's stringent claims are always claims on his *own* government, never on the governments of other countries, even if these claims are generated by the policy actions of foreign governments.

Consider now the issues that arise in formalizing what citizen i feels he is owed in a world of international trade. We are imagining that, if asked, citizen i could make a statement of the form "*I am owed a world in which '...blah...',*" where the phrase '*...blah...*' represents a set of conditions on the objects over which citizen i has moral sentiments that, when met, would in the eyes of citizen i describe a minimally fair counterfactual world with trade, the analog of the fair "reference transaction" described by Kahneman, Knetsch, and Thaler (1986). We allow these objects to represent both citizen i's distributive justice concerns and his procedural justice concerns.

For example, citizen i might state: "*I am owed a world in which* my *material utility is not pushed below that of my parents, and where I can live and work in the town where I grew up.*" Here, what citizen i feels he is owed can be expressed as a combination of two conditions, one relating to distributive justice concerns (*my material utility is not pushed below that of my parents*) and one relating to procedural justice concerns related to the way that material utility is delivered (*I can live and work in the town where I grew up*). Or citizen i might state: "*I am owed a world where I don't have to compete with – or buy goods that were produced with – the labor of foreign workers who are earning $2/hour or are working in unsafe conditions,*" two conditions that relate to procedural

[14] Fehr and Schmidt (1999) adopt a self-centered concept of fairness as it relates to inequality aversion. As they explain,

> Inequity aversion is self-centered if people do not care per se about inequity that exists among other people but are only interested in the fairness of their own material payoff relative to the payoff of others. (p 819).

In section 6 we consider the possibility of fairness concerns that extend to the treatment of others, either nationally or transnationally.

justice concerns (one about the workplace conditions in the foreign facilities that produce the goods with which citizen i's production must compete, and one about the workplace conditions in the foreign facilities that produce the goods that citizen i must consume).

Notice that citizen i's material utility might in fact be lower in the world he feels he is owed (e.g., if he could not buy goods that were produced with the labor of foreign workers who are working in unsafe conditions), but citizen i might still feel that he is owed such a world on moral grounds and could prefer that world (because he would "sleep better at night"). This possibility can arise if an individual has procedural justice concerns, while it cannot arise with distributive justice concerns only.

How can these diverse possibilities for what citizen i might feel he is owed in a world of international trade be formalized?

We assume that citizen i understands the model world economy in which he lives, and that his stringent claims are *coherent*, in the specific sense that the minimally fair counterfactual world that he conjures up to define his stringent claims is actually achievable given the policy instruments that the domestic and foreign governments have at their disposal. This assumption ensures that there will always exist sets of policies that would be deemed "fair" by citizen i in the sense that under those policies citizen i would receive everything he feels he is owed, both in terms of his distributive justice concerns and in terms of his procedural justice concerns.

If citizen i is concerned only about distributive justice, then formalizing citizen i's stringent claims is straightforward, as these claims amount to a number, namely, the level of material utility that citizen i would receive in the minimally fair counterfactual world that he conjures up to define his stringent claims.[15] But if citizen i has concerns for procedural justice, either in addition to or instead of concerns for distributive justice, it is no longer obvious how citizen i's stringent claims should be formalized.

Our approach is to assume that if citizen i has moral sentiments related to procedural justice concerns, then what he feels he is owed in relation to these concerns can always be converted into *a level of his material utility.*[16]

[15] In this case our aggrievement function can be related to the discussion of aggrievement in Passarelli and Tabellini (2017). They model aggrievement as arising when citizen i's material utility falls short of the material utility that citizen i would receive in a counterfactual world in which he believes he is treated fairly. This fair level of material utility, which is the analog of our stringent claims concept in the case where citizen i is concerned only about distributive justice, is determined in Passarelli and Tabellini by the policies that the government would choose if it placed what citizen i believes to be a fair weight on his utility when making its policy choice. As mentioned earlier, our focus is not on where fairness concerns come from, but rather on the effect of those concerns on the design of trade agreements. Therefore, in contrast to Passarelli and Tabellini, we take the determination of stringent claims as exogenous (and dictated by the sovereign right of citizen i to have his own moral sentiments), and instead we allow for the possibility that citizen i's notion of fairness extends beyond distributive justice concerns to include as well the possibility of procedural justice concerns, as we next describe in the text.

[16] This assumption may sound overly transactional, but recall that we have abstracted from altruistic motives and are focusing on a self-centered notion of fairness, and with this in

In particular, once citizen i has described the world he believes he is owed, we assume that he can then also express the level of material utility $V^{\mathcal{F}i}$ he would need to receive in order to be *fairly compensated* for living in a world characterized by any pattern of *deviations* from the procedural justice that his fair counterfactual world exhibits. And with this conversion, under our assumption we can express what citizen i feels he is owed in *any* hypothetical world – including a world where policies deviate in arbitrary ways from the sets of policies that would deliver procedural justice in the eyes of citizen i – as precisely this level of citizen i's material utility. In short, $V^{\mathcal{F}i}$ represents for domestic citizen i what in Risse's (2007) terminology would be called citizen i's *stringent claims*, a term that we will associate with $V^{\mathcal{F}i}$ in what follows.

In its most general form, citizen i's stringent claims can be written as a function of all of the domestic and foreign policies,

$$V^{\mathcal{F}i} = \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*), \tag{6}$$

where we can think of the arguments of $\hat{V}^{\mathcal{F}i}(\cdot)$ as representing the universe of possible concerns that citizen i could have regarding procedural justice.[17] When evaluated at any set of policies that satisfy citizen i's concerns over procedural justice, $V^{\mathcal{F}i}$ reflects only citizen i's distributive justice concerns, that is, what level of material utility citizen i feels he is owed when his procedural justice concerns have been met. This "baseline" level of material utility could be zero if citizen i has no distributive-justice-based stringent claims; or it could be a positive number if citizen i feels he is owed something on grounds of distributive justice (e.g., the same income level that his parents enjoyed). When evaluated at policy combinations that fail to meet all of citizen i's concerns over procedural justice, $V^{\mathcal{F}i}$ must then rise above this baseline level to a level that citizen i feels would fairly compensate him for living in a world where his procedural justice concerns are not met. For instance, individual i might think that he is owed a certain level of material utility if foreign workers are protected by strong safety standards, while he might think that he is owed a higher level of material utility

mind the assumption may be less objectionable as it essentially amounts to a rejection of lexicographic preferences over the various categories of moral sentiments (see Konow, 2003, pp 1234-1235, for a discussion of evidence from experimental and survey studies that seems to contradict the assumption of lexicographic preferences over competing moral sentiments). In fact, we would argue that for most considerations this assumption is relatively benign, as the level of material utility that citizen i might feel he is owed in a given situation could be extremely high, just not infinite. That said, for a consideration such as slavery our assumption taken literally would imply that there is some finite level of material utility where citizen i would feel that he received what he was owed and therefore was treated fairly even if he was enslaved. For this case we could allow the level of material utility to be arbitrarily high, but assigning it a value of infinity would reflect better what most people would have in mind for this situation. In this light, the domain of circumstances to which our formal analysis applies should be understood under the appropriate caveats.

[17] We do not include foreign transfers as an argument of $\hat{V}^{\mathcal{F}i}(\cdot)$, because under our assumption that preferences are identical and homothetic across citizens within each country and in the absence of transnational altruism, foreign transfers can have no bearing on domestic citizen i's stringent claims. We will return to this point in section 6, where we examine transnational fairness concerns.

if they are not. Based on our self-centered notion of fairness, this additional compensation is not motivated by altruistic concerns about the condition of foreign workers. Rather, it is driven by citizen i's perception that weak safety standards beyond the border distort competition in an unfair manner, and to avoid aggrievement at the prospect of living in such a world citizen i must then be compensated. Hence, $\hat{V}^{\mathcal{F}i}$ is weakly increasing (weakly decreasing) in any policy whose increase leads to a weak deterioration (weak improvement) in procedural justice in the eyes of citizen i.[18]

A key question is whether domestic citizen i's stringent claims as represented by $V^{\mathcal{F}i}$ actually depend on the foreign policies s^* and τ^*, and if so, what the nature of that dependence is. The answer to this question will be determined by the breadth and nature of citizen i's moral sentiments.

If citizen i's procedural justice concerns do not extend to the foreign policies s^* and τ^*, then (6) collapses to

$$V^{\mathcal{F}i} = \mathring{V}^{\mathcal{F}i}(s, T^i, \tau). \tag{7}$$

In this case, the range of domestic citizen i's moral sentiments beyond domestic distributive justice are restricted to procedural justice concerns within the domestic country – concerns about the levels of the three domestic policies s, T^i and τ that together dictate the domestic processes that generate citizen i's material utility – corresponding to the moral sentiments in the top left and top right boxes of our taxonomy in Figure 1.

On the other hand, if citizen i's procedural justice concerns do extend to the foreign policies s^* and τ^*, then there are two possibilities.

One possibility is that domestic citizen i's concerns for procedural justice are violated when the foreign government's *overall* policy stance – as reflected in the combined impact of s^* and τ^* on the position of the foreign export supply curve $E^*(s^*, p^*(\tau^*, p^w), p^w)$ – leads to a level of foreign trade pressure in the domestic economy – as summarized by the level of $\tilde{p}^w(s, \tau, s^*, \tau^*)$ – that citizen i deems to be unfair. If this characterizes the dependence of $V^{\mathcal{F}i}$ on s^* and τ^*, then we can take account of that structure and write

$$\begin{aligned} V^{\mathcal{F}i} &= \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*) = \tilde{V}^{\mathcal{F}i}(s, T^i, \tau, \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ &= \bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w)), \end{aligned} \tag{8}$$

where the second line of (8) follows from the last expression in the first line using the arbitrage condition (1) that $p = \tau p^w$.

As the form of domestic citizen i's stringent claims function in the second line of (8) makes clear, with the extra structure implied under this first possibility we can think of the range of domestic citizen i's moral sentiments beyond domestic distributive justice as again restricted to procedural justice concerns within the domestic country – this time, concerns about the levels of the three domestic policies s, T^i and τ *and* the domestic relative price $p(\tau, \tilde{p}^w)$ that together dictate the domestic processes that generate citizen i's material utility –

[18] In what follows we will also assume that each of the stringent claims functions we consider is continuously differentiable in each of its arguments.

and corresponding again to the moral sentiments in the top left and top right boxes of our taxonomy in Figure 1.

For example, citizen i might work in the domestic import-competing sector x and feel that it is unfair that a particularly low level of $p(\tau, \tilde{p}^w)$ – and hence holding fixed τ, a particularly low level of $\tilde{p}^w$ – is supported by an aggressively trade-promoting policy stance in the foreign country (e.g., a low foreign import tariff τ^* coupled with a weak foreign standard s^* in the foreign export sector). Or citizen i might work in the domestic export sector y and feel that it is unfair that a particularly high level of $p(\tau, \tilde{p}^w)$ – and hence holding fixed τ, a particularly high level of $\tilde{p}^w$ – is supported by an aggressively trade-protecting policy stance in the foreign country (e.g., a high foreign import tariff τ^* coupled with a high foreign standard s^* in the foreign export sector). Finally, notice also that the case where citizen i's procedural justice concerns do not extend to the foreign policies s^* and τ^* at all, which leads to the stringent claims function in (7), is a special case of (8), so we can put (7) to the side and without loss of generality focus on (8).[19]

The other possibility is that domestic citizen i's concerns for procedural justice are violated when a *particular* foreign policy is chosen, such as a weak foreign standard s^*, or a high foreign tariff τ^*. Put differently, under this possibility even holding fixed the level of foreign trade pressure $\tilde{p}^w(s, \tau, s^*, \tau^*)$, the level of material utility that citizen i feels he is owed would depend on the *specific mix of foreign policies* that lie behind this foreign trade pressure. In this case, the foreign policy itself is what is reprehensible in the eyes of domestic citizen i, and it is therefore the foreign policy itself that violates citizen i's sense of procedural justice. When this is the case, the special structure imposed in (8) is not valid, and instead (6) applies. And as (6) makes clear, under this second possibility we can think of the range of domestic citizen i's moral sentiments beyond domestic distributive justice as spanning procedural justice concerns everywhere, corresponding to the top left, top right and bottom right boxes of our taxonomy in Figure 1.[20]

[19] Having shown under this first possibility that domestic citizen i's stringent claims can be written as $\bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w))$, it can now also be seen that the fairness considerations that arise under this possibility are consistent with what Meyer (forthcoming, p 11) calls the traditional "production jurisdiction" view of international trade law, namely, a view that each country has the sole authority to tax and regulate the productive activities within its borders, and therefore "only the state in whose territory production occurs may tax or regulate a product or service *based on characteristics of its production*" (emphasis in the original). This suggests in turn that such fairness concerns might not create novel issues for a trade agreement to deal with, a suggestion that we show below is largely – but not completely – correct.

[20] The bottom left box of Figure 1 will become relevant in section 6 when we consider the possibility of (transnational) altruism. We can also relate this second possibility to the distinction drawn by Meyer (forthcoming) between production jurisdiction (see note 19) and what Meyer terms "consumption jurisdiction," whereby countries assert the authority to tax and regulate the production of the goods and services they consume based on characteristics of its production (e.g., the standards under which workers must work), regardless of where in the world that production takes place. Our stringent claims function $\hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*)$ for domestic citizen i is related to Meyer's notion of consumption jurisdiction, though these claims arise from moral sentiments only and may not depend on whether citizen i or anyone in his country is actually consuming the foreign good or service in question (e.g., domestic citizen

Aggrievement Armed with the stringent claims functions in (6) and (8), and using the expression for citizen i's material utility as defined in (5), we can now write down the aggrievement function for domestic citizen i. In particular, we assume that citizen i is not aggrieved as long as he receives in terms of his own material utility at least what he is owed, and that citizen i's level of aggrievement rises in the magnitude of any shortfall between what he is owed and what he receives.[21]

In the case where citizen i's stringent claims function takes the form in (8), his aggrievement function is given by

$$\begin{aligned} A^i &= A^i(\max[0, \bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w)) - V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))]) \\ &\equiv \mathcal{A}^i(\bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w)), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))), \end{aligned} \tag{9}$$

where $\mathcal{A}^i(\cdot)$ is weakly increasing in its first argument and weakly decreasing in its second argument. In the case where citizen i's stringent claims function takes the form in (6), his aggrievement function is given by

$$\begin{aligned} A^i &= A^i(\max[0, \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*) - V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))]) \\ &\equiv \mathcal{A}^i(\hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))), \end{aligned} \tag{10}$$

where again $\mathcal{A}^i(\cdot)$ is weakly increasing in its first argument and weakly decreasing in its second argument.[22]

Notice that the foreign policies s^* and τ^* appear in domestic citizen i's general aggrievement function as defined in (10) in two ways: through their impact on citizen i's stringent claims $\hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*)$, that is, through the level of material utility that citizen i feels he is owed, where they enter directly; and through their impact on citizen i's material utility $V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))$, where they enter only through the equilibrium relative world price $\tilde{p}^w(s, \tau, s^*, \tau^*)$. By contrast, when citizen i's moral sentiments satisfy the special structure that leads to the aggrievement function defined in (9), the foreign policies s^* and τ^* enter only through the equilibrium relative world price $\tilde{p}^w(s, \tau, s^*, \tau^*)$. As we will demonstrate below, this is the key structure that we exploit in our analysis of the purpose and design of fair trade agreements.

---

i may be aggrieved by the existence of unsafe working conditions in the foreign country, even if this is because he must compete with the foreign country for exports to a third country and neither he nor anyone in his country consumes any of the goods or services produced under those working conditions).

[21] We should note that according to Risse (2007), fairness does not require that stringent claims per se be satisfied, but only that they be satisfied in a proportional sense. As Risse explains,

> ... Suppose we are all owed a medication, and the more we take of it, the more we recover. No considerations other than medical need enter (disregard ownership, who is more deserving, and so on), and the needs are equal. Suppose there is not enough to restore everybody completely. Nobody can complain that her claim is not fully satisfied if all are satisfied equally. (Risse, 2007, p 358)

This feature is not incorporated into our aggrievement functions below, but it easily could be without changing our results and so we don't emphasize it in our discussion.

[22] We also assume that each of these aggrievement functions is continuously differentiable in each of its arguments.

Finally, we represent domestic citizen i's total welfare, denoted by W^i, by his material utility minus the aggrievement that he suffers. And while above we have taken the perspective of the domestic country, a completely analogous set of derivations holds for the foreign country, leading to an analogous representation of foreign citizen i's material utility, aggrievement function, and total welfare, denoted respectively by V^{*i}, $\mathcal{A}^{*i}$ and W^{*i}.

For future reference, we will refer to the case where stringent claims are restricted to own-country distributive and procedural justice concerns – and therefore take the form in (9) – as *Case I*, and we will refer to the case where stringent claims include own-country distributive justice concerns and procedural justice concerns everywhere – and therefore take the form in (10) – as *Case II*.[23] We record here the total welfare functions for domestic and foreign citizen i for each of these two cases:[24]

Case I: Own-Country Distributive and Procedural Justice

$$\begin{aligned} W^i &= V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ &\quad - \mathcal{A}^i(\bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w)), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ &\quad \equiv \bar{W}^i(V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w))) \end{aligned} \tag{11}$$

$$\begin{aligned} &W^{*i} = V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ &- \mathcal{A}^{*i}(\bar{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, p^*(\tau^*, \tilde{p}^w)), V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ &\equiv \bar{W}^{*i}(V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, p^*(\tau^*, \tilde{p}^w))) \end{aligned} \tag{12}$$

with $\bar{W}^i$ and $\bar{W}^{*i}$ increasing in their first arguments and weakly decreasing in their second arguments.

[23] To avoid an unnecessary taxonomy, we do not consider here the mixed cases where the stringent claims of some citizens take the form in (9) while the stringent claims of some other citizens take the form in (10), but such cases can be understood from the two cases we consider below (and we allow for them in the heterogeneous agent model of section 5).

[24] As we noted earlier in the context of an example where domestic citizen i has moral sentiments about consuming goods that are produced under unsafe conditions in the foreign country, whether citizen i would actually prefer to live in a world in which he buys goods produced with foreign workers who work under unsafe conditions would depend on the strength of his moral sentiments. This can now be seen in light of the total welfare of citizen i as recorded below. Consider for example Case II. If citizen i's material welfare V^i is sufficiently increased by the ability to consume cheap foreign goods that are produced in unsafe conditions, and if his moral sentiment about consuming such goods is relatively weak so that $\hat{V}^{\mathcal{F}i}$ is not too high in a world where he consumes these goods, then citizen i's aggrievement $\mathcal{A}^i$ from living in such a world could be small or even zero; and his total welfare W^i could then be higher in such a world and he would prefer to live in that world.

Case II: Own-Country Distributive Justice and Procedural Justice Everywhere

$$\begin{aligned} W^i &= V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ &\quad - \mathcal{A}^i(\hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ &\quad \equiv \hat{W}^i(V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*)) \end{aligned} \quad (13)$$

$$\begin{aligned} W^{*i} &= V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ &\quad - \mathcal{A}^{*i}(\hat{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, s, \tau), V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ &\equiv \hat{W}^{*i}(V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, s, \tau)) \end{aligned} \quad (14)$$

with $\hat{W}^i$ and $\hat{W}^{*i}$ increasing in their first arguments and weakly decreasing in their second arguments.

In sections 4 and 5 we consider the purpose of a trade agreement in the presence of fairness considerations, where these fairness consideration can either take the form described by Case I or Case II. In section 4 we consider a representative agent version of the model presented above, where distributive justice issues are constrained to be transnational (e.g., the representative citizen in the domestic country feels he is owed an equal division of the gains from trade with the foreign country). We then turn in section 5 to consider the version of this model with heterogeneous agents, where within-country distributive justice issues may also arise and where a preference for predistribution over redistribution can therefore be considered. In each case, we proceed as follows. First we ask how concerns for fairness impact the internationally efficient policy choices, where efficiency is judged relative to the objective functions of the two governments. Then we ask how concerns for fairness impact noncooperative Nash policy choices. And finally, we follow the approach of Bagwell and Staiger (1999, 2001, 2002) and ask what accounts for the difference between the Nash and internationally efficient policy choices in this setting, and with this account we then identify the purpose of a trade agreement in the presence of fairness considerations and draw conclusions about its appropriate design.

4 A Representative Agent

We begin with a representative-agent version of the model presented in section 3, where distributive justice issues are constrained to be transnational. In addition to suppressing the citizen subscript i, we can also omit the government transfer policies and set $T \equiv 0 \equiv T^*$, since such transfers have no role in this setting.

4.1 Case I: Own-Country Distributive and Procedural Justice

When the representative agent in each country has own-country distributive and procedural justice concerns corresponding to the moral sentiments in the

top left and top right boxes of Figure 1, we can write the total welfare functions of the domestic and foreign citizen given in (11) and (12) respectively as

$$\begin{aligned} W = V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ - \mathcal{A}(\bar{V}^{\mathcal{F}}(s, \tau, p(\tau, \tilde{p}^w)), V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ \equiv \bar{W}(V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}}(s, \tau, p(\tau, \tilde{p}^w))) \end{aligned} \quad (15)$$

$$\begin{aligned} W^* = V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ - \mathcal{A}^*(\bar{V}^{\mathcal{F}*}(s^*, \tau^*, p^*(\tau^*, \tilde{p}^w)), V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ \equiv \bar{W}^*(V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}*}(s^*, \tau^*, p^*(\tau^*, \tilde{p}^w))) \end{aligned} \quad (16)$$

with $\bar{W}$ and $\bar{W}^*$ increasing in their first arguments and weakly decreasing in their second arguments.

To characterize the domestic and foreign policy choices, we now introduce the objectives of the domestic and foreign government, which we denote by G and G^* respectively. We remain agnostic about the weight that a government gives to its citizen's aggrievement, and we interpret a government's dislike of its citizen's aggrievement either as reflecting its willingness to take care of their psychological well-being, or as reflecting a dislike of the (unmodeled) costs that its aggrieved citizen imposes on the government (e.g., protests, shading behavior). To capture this in a simple way, we assume that each government chooses its policy instruments with the goal of maximizing the material utility of its representative citizen minus a weight ($\gamma \geq 0$ for the domestic government, $\gamma^* \geq 0$ for the foreign government) times the aggrievement suffered by its representative citizen. Formally we specify G and G^* as follows:

$$\begin{aligned} G = V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ - \gamma \times \mathcal{A}(\bar{V}^{\mathcal{F}}(s, \tau, p(\tau, \tilde{p}^w)), V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ \equiv \bar{G}(V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}}(s, \tau, p(\tau, \tilde{p}^w)), \gamma) \end{aligned} \quad (17)$$

$$\begin{aligned} G^* = V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ - \gamma^* \times \mathcal{A}^*(\bar{V}^{\mathcal{F}*}(s^*, \tau^*, p^*(\tau^*, \tilde{p}^w)), V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ \equiv \bar{G}^*(V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \bar{V}^{\mathcal{F}*}(s^*, \tau^*, p^*(\tau^*, \tilde{p}^w)), \gamma^*) \end{aligned} \quad (18)$$

with $\bar{G}$ and $\bar{G}^*$ increasing in their first arguments and weakly decreasing in their second arguments. We will focus on the case where $\gamma > 0$ and $\gamma^* > 0$, which represents the case where the governments are responsive to fairness considerations of their citizens, and which for shorthand we will refer to as the case where fairness concerns are "present"; when $\gamma = 0 = \gamma^*$ and fairness considerations are absent, our model collapses to (in this section a representative-agent version of) the model of Bagwell and Staiger (2001), with all of the results implied therein.

Notice that $\tilde{p}^w$ enters $\bar{G}$ and $\bar{G}^*$ directly only through the representative citizen's material welfare V, which by Assumption 1 is decreasing in $\tilde{p}^w$ for the domestic representative citizen and by the analogous assumption for the foreign country is increasing in $\tilde{p}^w$ for the foreign representative citizen. Hence using also the derivative properties of the $\bar{G}$ and $\bar{G}^*$ functions we have:

$$\frac{\partial \bar{G}(V(s,p(\tau,\tilde{p}^w),\tilde{p}^w(s,\tau,s^*,\tau^*)),\bar{V}^{\mathcal{F}}(s,\tau,p(\tau,\tilde{p}^w)),\gamma)}{\partial \tilde{p}^w} = \bar{G}_V V_{p^w} < 0$$

$$\frac{\partial \bar{G}^*(V^*(s^*,p^*(\tau^*,\tilde{p}^w),\tilde{p}^w(s,\tau,s^*,\tau^*)),\bar{V}^{\mathcal{F}*}(s^*,\tau^*,p^*(\tau^*,\tilde{p}^w)),\gamma^*)}{\partial \tilde{p}^w} = \bar{G}^*_{V^*} V^*_{p^w} > 0,$$

where we use subscripts to denote partial derivatives.

Internationally efficient policies We are now ready to characterize the internationally efficient choices of the domestic and foreign policies, which we denote by s^E, τ^E, s^{*E} and τ^{*E}. Recalling that we define efficiency with respect to the government objectives $\tilde{G}$ and $\tilde{G}^*$, these policies solve the following program:

$$\begin{aligned} \max_{s,\tau,s^*,\tau^*} \quad & \bar{G}(V(s,p(\tau,\tilde{p}^w),\tilde{p}^w(s,\tau,s^*,\tau^*)),\bar{V}^{\mathcal{F}}(s,\tau,p(\tau,\tilde{p}^w)),\gamma) \\ & s.t. \\ \bar{G}^*(V^*(s^*,p^*(\tau^*,\tilde{p}^w),\tilde{p}^w(s,\tau,s^*,\tau^*)),\bar{V}^{\mathcal{F}*}(s^*,\tau^*,p^*(\tau^*,\tilde{p}^w)),\gamma^*) &\geq \bar{G}^{*E} \end{aligned} \tag{19}$$

where $\bar{G}^{*E} \equiv \bar{G}^*(V^*(s^{*E},p^*(\tau^{*E},\tilde{p}^{wE}),\tilde{p}^{wE}),\bar{V}^{\mathcal{F}*}(s^{*E},\tau^{*E},p^*(\tau^{*E},\tilde{p}^{wE})),\gamma^*)$ and $\tilde{p}^{wE} \equiv \tilde{p}^w(s^E,\tau^E,s^{*E},\tau^{*E})$.

Forming the Lagrangian associated with (19) and manipulating the first order conditions to eliminate the Lagrange multiplier yields the three conditions that define the international efficiency frontier:[25]

$$\begin{aligned} \left[\bar{G}_V V_p + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}^{\mathcal{F}}_p\right] \times \tilde{p}^w + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}^{\mathcal{F}}_\tau + \\ \left[\bar{G}_V V_s + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}^{\mathcal{F}}_s\right] \times \frac{ds}{d\tau}|_{d\tilde{p}^w=0} = 0 \end{aligned} \tag{20}$$

$$\begin{aligned} \left[\bar{G}^*_{V^*} V^*_{p^*} + \bar{G}^*_{\bar{V}^{\mathcal{F}*}} \bar{V}^{\mathcal{F}*}_{p^*}\right] \times \left(-\frac{p^*}{\tau^*}\right) + \bar{G}^*_{\bar{V}^{\mathcal{F}*}} \bar{V}^{\mathcal{F}*}_{\tau^*} + \\ \left[\bar{G}^*_{V^*} V^*_{s^*} + \bar{G}^*_{\bar{V}^{\mathcal{F}*}} \bar{V}^{\mathcal{F}*}_{s^*}\right] \times \frac{ds^*}{d\tau^*}|_{d\tilde{p}^w=0} = 0 \end{aligned} \tag{21}$$

[25] We assume that the curvature properties of the functions we have defined ensure that the second-order conditions are satisfied for all maximization problems we consider below.

$$\left[\frac{\bar{G}_V\left[V_p\frac{dp}{d\tau}+V_{p^w}\frac{\partial\tilde{p}^w}{\partial\tau}\right]+\bar{G}_{\bar{V}^{\mathcal{F}}}\left(\bar{V}_\tau^{\mathcal{F}}+\bar{V}_p^{\mathcal{F}}\frac{dp}{d\tau}\right)}{\left[\frac{1}{\tau^*}\left(\bar{G}^*_{V^*}V^*_{p^*}+\bar{G}^*_{\bar{V}^{\mathcal{F}_*}}\bar{V}^{\mathcal{F}*}_{p^*}\right)+\bar{G}^*_{V^*}V^*_{p^w}\right]\frac{\partial\tilde{p}^w}{\partial\tau}}\right]=$$

$$\left[\frac{\left[\tau\left(\bar{G}_V V_p+\bar{G}_{\bar{V}^{\mathcal{F}}}\bar{V}_p^{\mathcal{F}}\right)+\bar{G}_V V_{p^w}\right]\frac{\partial\tilde{p}^w}{\partial\tau^*}}{\bar{G}^*_{V^*}\left[V^*_{p^*}\frac{dp^*}{d\tau^*}+V^*_{p^w}\frac{\partial\tilde{p}^w}{\partial\tau^*}\right]+\bar{G}^*_{\bar{V}^{\mathcal{F}_*}}\left(\bar{V}^{\mathcal{F}*}_{\tau^*}+\bar{V}^{\mathcal{F}*}_{p^*}\frac{dp^*}{d\tau^*}\right)}\right] \quad (22)$$

where in writing these conditions we have used the fact that

$$\frac{ds}{d\tau}|_{d\tilde{p}^w=0}=-\left(\frac{\partial d\tilde{p}^w/\partial\tau}{\partial d\tilde{p}^w/\partial s}\right) \quad \text{and} \quad \frac{ds^*}{d\tau^*}|_{d\tilde{p}^w=0}=-\left(\frac{\partial d\tilde{p}^w/\partial\tau^*}{\partial d\tilde{p}^w/\partial s^*}\right).$$

As in Bagwell and Staiger (2001), the conditions in (20) and (21) can be interpreted as "national" efficiency conditions for the domestic and foreign country, respectively. Condition (20) says that at internationally efficient policy choices, the domestic government should be indifferent to a small increase in τ combined with a small change in s that holds the equilibrium relative world price $\tilde{p}^w$ fixed. This is because by holding $\tilde{p}^w$ fixed, such domestic-country policy changes do not impact the foreign government, as inspection of the expression for $\bar{G}^*$ in (18) confirms, and so international efficiency dictates that the domestic government must be indifferent to these policy changes as well. The key structure reflected in (20) is that it isolates a condition for international efficiency that *only involves tradeoffs as perceived by the domestic government*. A similar interpretation applies for condition (21) as it relates to the foreign government choices of τ^* and s^*. The condition in (22) can then be interpreted as the "international" efficiency condition because, in combination with (20) and (21), condition (22) determines the levels of τ and τ^* that generate the efficient trade volumes between the two countries.

Non-cooperative Nash policies We next characterize the noncooperative (Nash) policy choices of the two governments. These are defined by the four first-order conditions:

$$\begin{aligned}
\bar{G}_V\left(V_p\frac{dp}{d\tau}+V_{p^w}\frac{\partial\tilde{p}^w}{\partial\tau}\right)+\bar{G}_{\bar{V}^{\mathcal{F}}}\left(\bar{V}_\tau^{\mathcal{F}}+\bar{V}_p^{\mathcal{F}}\frac{dp}{d\tau}\right) &= 0\\
\bar{G}_V\left(V_s+\left[\tau V_p+V_{p^w}\right]\frac{\partial\tilde{p}^w}{\partial s}\right)+\bar{G}_{\bar{V}^{\mathcal{F}}}\left(\bar{V}_s^{\mathcal{F}}+\tau\bar{V}_p^{\mathcal{F}}\frac{\partial\tilde{p}^w}{\partial s}\right) &= 0 \quad (23)\\
\bar{G}^*_{V*}\left(V^*_{p^*}\frac{dp^*}{d\tau^*}+V^*_{p^w}\frac{\partial\tilde{p}^w}{\partial\tau^*}\right)+\bar{G}^*_{\bar{V}^{\mathcal{F}_*}}\left(\bar{V}^{\mathcal{F}*}_{\tau^*}+\bar{V}^{\mathcal{F}*}_{p^*}\frac{dp^*}{d\tau^*}\right) &= 0\\
\bar{G}^*_{V^*}\left(V^*_{s^*}+\left[\frac{1}{\tau^*}V^*_{p^*}+V^*_{p^w}\right]\frac{\partial\tilde{p}^w}{\partial s^*}\right)+\bar{G}^*_{\bar{V}^{\mathcal{F}_*}}\left(\bar{V}^{\mathcal{F}*}_{s^*}+\frac{1}{\tau^*}\bar{V}^{\mathcal{F}*}_{p^*}\frac{\partial\tilde{p}^w}{\partial s^*}\right) &= 0.
\end{aligned}$$

Comparing the Nash policies that must satisfy (23) to the internationally efficient policies characterized by (20)-(22), it is direct to show that the top two

conditions in (23) together imply that the domestic national efficiency condition (20) is satisfied in the Nash equilibrium; this is intuitive, since the domestic national efficiency condition (20) says that the domestic government should be indifferent to a small increase in τ combined with a small change in s that holds the equilibrium relative world price $\tilde{p}^w$ fixed, and the top two Nash first-order conditions in (23) ensure that the domestic government will be indifferent to *any* small change in τ or s. Similarly it is direct to show that the bottom two conditions in (23) together imply that the foreign national efficiency condition (21) is satisfied in the Nash equilibrium as well. But the first and third conditions in (23) together imply that the international efficiency condition (22) is violated in the Nash equilibrium. Hence, as Bagwell and Staiger (2001) conclude in the absence of fairness considerations, we may conclude that when fairness considerations regarding own-country distributive and procedural justice concerns are present as in Case I, Nash policy choices are internationally inefficient for a single reason, namely, because of the inefficient trade volumes they imply.

Shallow integration It is now also possible to see that even when fairness considerations are present, a GATT-like shallow approach to efficient integration remains feasible as long as these considerations conform to Case I and are therefore limited to own-country distributive and procedural justice concerns. In effect, with each government in the Nash equilibrium choosing an efficient mix of its own standards and tariffs that nevertheless together generate inefficient trade volumes, governments can focus on negotiating tariff levels that would imply efficient trade volumes and therefore satisfy the international efficiency condition (22) in light of their Nash standards. And with these trade volumes implying a level of the equilibrium relative world price $\tilde{p}^w$, each government can then be allowed to adjust its mix of standards and tariffs unilaterally as long as its adjustments do not alter the level of $\tilde{p}^w$, adjustments which will ensure that the domestic and foreign national efficiency conditions (20) and (21) are then satisfied as well. As Bagwell and Staiger (2001, 2002) observe, this procedure conforms well with the GATT/WTO tradition of tariff-led "market access" negotiations, under which negotiated tariff bindings imply market access commitments that are protected from erosion with various GATT/WTO Articles that govern permissible non-tariff policy interventions.

Following Bagwell and Staiger (2001, 2002), this last point can be formalized by defining *market access* as the volume of imports a country would accept at a particular world price, a definition which links the concept of market access to the position of a country's import demand curve.[26] For example, the domestic-country market access at the world price $\hat{p}^w$ implied by domestic policies τ and s would be given by $M(s, p(\tau, \hat{p}^w), \hat{p}^w)$, and the foreign-country market access at the world price $\hat{p}^w$ implied by foreign policies τ^* and s^* would be given by $M^*(s^*, p^*(\tau^*, \hat{p}^w), \hat{p}^w)$, where M^* denotes foreign-country imports of good

[26] The link between the position of a country's import demand curve and the definition of market access within the GATT/WTO – "the competitive relationship between imported and domestic products" – was first proposed by Bagwell and Staiger (2001), and is described more fully in Bagwell and Staiger (2002, pp 29-30).

y and $M^*(s^*, p^*(\tau^*, \hat{p}^w), \hat{p}^w) = \hat{p}^w E^*(s^*, p^*(\tau^*, \hat{p}^w), \hat{p}^w)$ by the foreign-country balanced trade condition which must hold for any world price $\hat{p}^w$. It is direct to confirm from (3) that changes in τ and s that do not alter domestic market access $M(s, p(\tau, \tilde{p}^w), \tilde{p}^w)$ evaluated at the market clearing world price $\tilde{p}^w$ cannot alter $\tilde{p}^w$; and similarly changes in τ^* and s^* that do not alter foreign market access $M^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w)$ evaluated at the market clearing world price $\tilde{p}^w$ (and hence do not alter $E^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w)$) cannot alter $\tilde{p}^w$.

Therefore, under a shallow integration approach, in a first step governments can focus their negotiations on achieving tariffs that in light of their Nash standards imply market access levels that together induce efficient trade volumes satisfying the international efficiency condition (22). And with these trade volumes implying a level of the equilibrium world price $\tilde{p}^w$, in a second step each government can then be allowed to make unilateral adjustments to its mix of standards and tariffs subsequent to the tariff negotiations, as long as its adjustments do not alter the level of market access that it committed to in the tariff negotiations and hence do not alter the market clearing world price $\tilde{p}^w$. This second step ensures that the domestic and foreign national efficiency conditions (20) and (21) are then satisfied as well. We summarize these points with:

Proposition 1 *When fairness considerations involving own-country distributive and procedural justice concerns are present, Nash policy choices are internationally inefficient for a single reason, namely, because of the inefficient equilibrium trade volumes they imply. In this setting, a GATT-like shallow approach to efficient integration remains feasible.*

Market access and the purpose of a trade agreement With market access defined, we may also ask whether it remains true in the presence of such fairness considerations that each government can gain from tariff negotiations only if it secures additional market access from its trading partner, as Bagwell and Staiger (2001) demonstrate is the case when fairness considerations are absent. To answer this question, we first note that, if the domestic government did not secure through negotiation additional market access from its trading partner for at least some world price, then as a result of these negotiations the foreign export supply curve would shift in (weakly), and under standard Marshall-Lerner stability conditions the negotiated foreign government policy changes would contribute toward a rise (weakly) in the equilibrium world price $\tilde{p}^w$. We wish to determine whether the domestic government could ever benefit from negotiations that involved foreign government policy changes of this kind. We confirm that this is in fact possible, but only if there is a moral "taste for free trade" in the domestic country.

To this end, we first write down the impact of changes in τ^* and s^* on domestic government welfare:

$$\frac{d\bar{G}}{d\tau^*} + \frac{d\bar{G}}{ds^*} = \left[\bar{G}_V \left(\tau V_p + V_{p^w}\right) + \bar{G}_{\bar{V}^{\mathcal{F}}} \tau \bar{V}_p^{\mathcal{F}}\right] \times \left(\frac{\partial \tilde{p}^w}{\partial \tau^*} + \frac{\partial \tilde{p}^w}{\partial s^*}\right).$$

Evaluating this impact when the domestic government is on its reaction curves as defined by the top two conditions in (23) yields

$$\frac{d\bar{G}}{d\tau^*}+\frac{d\bar{G}}{ds^*} = \left[\left(\bar{G}_V\left(1-\theta\tau^R(s^*,\tau^*)\right)V_{p^w}\right)+\left(-\bar{G}_{\bar{V}^{\mathcal{F}}}\frac{\tau^R(s^*,\tau^*)\bar{V}_\tau^{\mathcal{F}}}{dp/d\tau}\right)\right]\times\left(\frac{\partial\tilde{p}^w}{\partial\tau^*}+\frac{\partial\tilde{p}^w}{\partial s^*}\right) \tag{24}$$

where $\tau^R(s^*,\tau^*)$ is the domestic government's reaction-curve tariff and where $\theta \equiv \frac{\partial\tilde{p}^w/\partial\tau}{d\tilde{p}^w/d\tau} < 0$. We can consider two cases.

First, suppose that $\bar{V}_\tau^{\mathcal{F}} \leq 0$ when evaluated at $\tau^R(s^*,\tau^*)$ for the relevant range of s^* and τ^*. If $\bar{V}_\tau^{\mathcal{F}} < 0$, then the domestic citizen has "protectionist" moral sentiments and feels he is owed a world in which the domestic tariff remains above some minimal level, while if $\bar{V}_\tau^{\mathcal{F}} = 0$ the domestic citizen feels he is owed nothing concerning the level of the domestic tariff per se. With $V_{p^w} < 0$ under Assumption 1, it then follows from the derivative properties of $\bar{G}$ that the term in square brackets in (24) is the sum of two negative terms, and hence if the domestic government can remain on its reaction curve it will be (weakly) hurt by any changes to s^* and τ^* that fail to expand foreign market access for at least some world price (and hence fail to contribute to a fall in $\tilde{p}^w$). And if the negotiations require the domestic government to move off its reaction curve, this can only be worse for the domestic government. Hence, provided that $\bar{V}_\tau^{\mathcal{F}} \leq 0$ when evaluated at $\tau^R(s^*,\tau^*)$ for the relevant range of s^* and τ^*, we may conclude that the domestic government can gain from tariff negotiations only if it secures additional market access from the foreign government (with an analogous argument applying to the foreign government), just as Bagwell and Staiger (2001) demonstrate is the case when fairness considerations are absent.

However, if $\bar{V}_\tau^{\mathcal{F}} > 0$ when evaluated at $\tau^R(s^*,\tau^*)$ for some s^* and τ^* in the relevant range, this result is no longer guaranteed. To see why, note that when $\bar{V}_\tau^{\mathcal{F}} > 0$, the second term in the square brackets in (24) will be positive; and if the magnitude of this second term is sufficiently large, it could dominate the first (negative) term in the square brackets and cause the overall sign of the term in square brackets in (24) to become positive. When this is the case, the result above would flip, and the domestic government could only gain from negotiations with the foreign government if the negotiated foreign policy changes led to a *reduction* in foreign-country market access for at least some world price (so that the negotiated foreign policy changes contributed to a *rise* in $\tilde{p}^w$). The case where $\bar{V}_\tau^{\mathcal{F}} > 0$ reflects a situation where the domestic citizen has "free trade" moral sentiments and feels he is owed a world in which the domestic tariff does not rise above some maximal level. To the extent that this second case is relevant, it points to the possibility that the properties of $\bar{V}_\tau^{\mathcal{F}}$ could have important implications for the design of a trade agreement.

In particular, this case raises the possibility that countries might, through their trade agreements, seek commitments from their trading partners to *restrict* trade volumes rather than expand them. On the one hand, this is something that the GATT/WTO mechanism of tariff *bindings* – the legal *maximum* for an applied tariff that a country commits to under GATT/WTO market access negotiations – would be ill-equipped to deliver. On the other hand, restric-

tions on export-promoting subsidies, such as those that were introduced into the GATT/WTO system under the 1995 WTO Agreement on Subsidies and Countervailing Measures, would be consistent with this desire when fairness concerns take the form of $\bar{V}_{\tau}^{\mathcal{F}} > 0$ for domestic citizens and $\bar{V}_{\tau^*}^{\mathcal{F}*} > 0$ for foreign citizens, and this is a desire that is difficult to explain when fairness concerns are absent (see, for example, Bagwell and Staiger, 2012).

Intuitively, the case where $\bar{V}_{\tau}^{\mathcal{F}} > 0$ and $\bar{V}_{\tau^*}^{\mathcal{F}*} > 0$ reflects a situation where citizens of each country feel that they are owed a world with low tariffs, and where each government's unilateral policy choices fail to internalize an international externality that takes a novel form, namely, that on the margin, a small reduction in the trade-promoting stance of one country's policies could allow the other country to adopt a slightly lower tariff and therefore better meet the stringent claims of its representative citizen while preserving the same material outcomes associated with its standards and local relative price.

This finding also foreshadows the role of $\bar{V}_{\tau}^{\mathcal{F}}$ in determining the purpose of a trade agreement more generally. This can be seen by asking a final question, namely, whether in the presence of fairness considerations that correspond to Case I, terms-of-trade manipulation and the attendant international cost-shifting incentive continues to be *the* problem that prevents Nash policy choices from reaching the efficiency frontier, as Bagwell and Staiger (2001) demonstrate is the case when fairness considerations are absent. To answer this question, we follow Bagwell and Staiger and define the *politically optimal* policies as those policies that would be chosen in the Nash equilibrium if both governments did not value their ability to manipulate the terms of trade and shift some of the costs of their policy intervention onto the other country, so that $V_{p^w} \equiv 0 \equiv V_{p^w}^*$. Using (23), the politically optimal policies are defined by

$$\begin{aligned}
\left(\bar{G}_V V_p + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}_p^{\mathcal{F}}\right) \frac{dp}{d\tau} + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}_{\tau}^{\mathcal{F}} &= 0 \\
\bar{G}_V \left(V_s + \tau V_p \frac{\partial \tilde{p}^w}{\partial s}\right) + \bar{G}_{\bar{V}^{\mathcal{F}}} \left(\bar{V}_s^{\mathcal{F}} + \tau \bar{V}_p^{\mathcal{F}} \frac{\partial \tilde{p}^w}{\partial s}\right) &= 0 \qquad (25) \\
\left(\bar{G}_{V*}^* V_{p^*}^* + \bar{G}_{\bar{V}^{\mathcal{F}*}}^* \bar{V}_{p^*}^{\mathcal{F}*}\right) \frac{dp^*}{d\tau^*} + \bar{G}_{\bar{V}^{\mathcal{F}*}}^* \bar{V}_{\tau^*}^{\mathcal{F}*} &= 0 \\
\bar{G}_{V^*}^* \left(V_{s^*}^* + \frac{1}{\tau^*} V_{p^*}^* \frac{\partial \tilde{p}^w}{\partial s^*}\right) + \bar{G}_{\bar{V}^{\mathcal{F}*}}^* \left(\bar{V}_{s^*}^{\mathcal{F}*} + \frac{1}{\tau^*} \bar{V}_{p^*}^{\mathcal{F}*} \frac{\partial \tilde{p}^w}{\partial s^*}\right) &= 0.
\end{aligned}$$

Our question now is whether the politically optimal policies defined by (25) satisfy the conditions for international efficiency defined by (20)-(22). If so, then we can conclude that the purpose of a trade agreement in the presence of fairness considerations that correspond to Case I is to eliminate terms-of-trade manipulation and the attendant international cost-shifting incentive from the unilateral policy choices of governments, just as Bagwell and Staiger (2001) argue is the case in the absence of fairness considerations. To highlight the role played by $\bar{V}_{\tau}^{\mathcal{F}}$ in the answer to this question, it is instructive to consider first the case where $\bar{V}_{\tau}^{\mathcal{F}} = 0 = \bar{V}_{\tau^*}^{\mathcal{F}*}$ when evaluated at politically optimal policy choices, and then the case where instead $\bar{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \bar{V}_{\tau^*}^{\mathcal{F}*}$.

When $\bar{V}_{\tau}^{\mathcal{F}} = 0 = \bar{V}_{\tau^*}^{\mathcal{F}*}$ at the politically optimal policy choices defined by the four conditions in (25), these conditions reduce to

$$\begin{aligned}
\left(\bar{G}_V V_p + \bar{G}_{\bar{V}^{\mathcal{F}}} \bar{V}_p^{\mathcal{F}}\right) \frac{dp}{d\tau} &= 0 \\
\bar{G}_V \left(V_s + \tau V_p \frac{\partial \tilde{p}^w}{\partial s}\right) + \bar{G}_{\bar{V}^{\mathcal{F}}} \left(\bar{V}_s^{\mathcal{F}} + \tau \bar{V}_p^{\mathcal{F}} \frac{\partial \tilde{p}^w}{\partial s}\right) &= 0 \\
\left(\bar{G}^*_{V^*} V^*_{p^*} + \bar{G}^*_{\bar{V}^{\mathcal{F}*}} \bar{V}^{\mathcal{F}*}_{p^*}\right) \frac{dp^*}{d\tau^*} &= 0 \\
\bar{G}^*_{V^*} \left(V^*_{s^*} + \frac{1}{\tau^*} V^*_{p^*} \frac{\partial \tilde{p}^w}{\partial s^*}\right) + \bar{G}^*_{\bar{V}^{\mathcal{F}*}} \left(\bar{V}^{\mathcal{F}*}_{s^*} + \frac{1}{\tau^*} \bar{V}^{\mathcal{F}*}_{p^*} \frac{\partial \tilde{p}^w}{\partial s^*}\right) &= 0.
\end{aligned} \tag{26}$$

But it is now direct to confirm that, when evaluated at the policies satisfying (26) along with the additional condition that $\bar{V}_{\tau}^{\mathcal{F}} = 0 = \bar{V}_{\tau^*}^{\mathcal{F}*}$ when evaluated at these policies, the three conditions in (20)-(22) that define the international efficiency frontier are satisfied. Hence, when fairness considerations involving own-country distributive and procedural justice concerns are present but with the additional condition that $\bar{V}_{\tau}^{\mathcal{F}} = 0 = \bar{V}_{\tau^*}^{\mathcal{F}*}$ when evaluated at politically optimal policies, we can conclude that the purpose of a trade agreement is to eliminate terms-of-trade manipulation and the attendant international cost-shifting incentive from the unilateral policy choices of governments, just as Bagwell and Staiger (2001) argue it is in the absence of fairness considerations.

Now consider the case where $\bar{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \bar{V}_{\tau^*}^{\mathcal{F}*}$ when evaluated at the politically optimal policies defined by (25). Evaluating the efficiency properties of these policies using the conditions for efficiency in (20)-(22), it can be confirmed that when $\bar{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \bar{V}_{\tau^*}^{\mathcal{F}*}$, politically optimal policies satisfy the two national efficiency conditions (20) and (21); but the international efficiency condition (22) is violated. Hence, when fairness considerations involving own-country distributive and procedural justice concerns are present and when it is also the case that $\bar{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \bar{V}_{\tau^*}^{\mathcal{F}*}$ at the politically optimal policies, so that at politically optimal policies citizens have stringent claims over their own-country tariff levels, the purpose of a trade agreement *cannot* be characterized as simply eliminating terms-of-trade manipulation and the attendant international cost-shifting incentive from the unilateral policy choices of governments: there is more to it than that. This helps explain why, when $\bar{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \bar{V}_{\tau^*}^{\mathcal{F}*}$, the design of a trade agreement may under some circumstances (i.e., when $\bar{V}_{\tau}^{\mathcal{F}} > 0$ and/or $\bar{V}_{\tau^*}^{\mathcal{F}*} > 0$) need to reflect a desire to negotiate *limits* on market access rather than an expansion of market access relative to Nash levels, as we have noted above.[27]

We summarize these points with:

[27] We noted that the arbitrage condition (1) allows the domestic stringent claims function $\tilde{V}^{\mathcal{F}}(s, \tau, \tilde{p}^w(s, \tau, s^*, \tau^*))$ in (8) to be written in the equivalent form $\bar{V}^{\mathcal{F}}(s, \tau, p(\tau, \tilde{p}^w))$. When stringent claims are instead written as $\tilde{V}^{\mathcal{F}}(s, \tau, \tilde{p}^w(s, \tau, s^*, \tau^*))$, there is an interpretation of the political optimum that corresponds to a point on the efficiency frontier even when $\tilde{V}_{\tau}^{\mathcal{F}} \neq 0 \neq \tilde{V}_{\tau^*}^{\mathcal{F}*}$, namely, when politically optimal policies are interpreted as those policies that would be chosen in the Nash equilibrium if both governments did not value their ability to manipulate the terms of trade in the sense that $V_{p^w} \equiv 0 \equiv V^*_{p^w}$ *and* $\tilde{V}^{\mathcal{F}}_{p^w} \equiv 0 \equiv \tilde{V}^{\mathcal{F}*}_{p^w}$.

Proposition 2 *When fairness considerations involving own-country distributive and procedural justice concerns are present, the purpose of a trade agreement hinges on whether stringent claims extend to the level of own-country tariffs. In particular, the purpose of a trade agreement is to eliminate terms-of-trade manipulation and the attendant international cost-shifting incentive from the unilateral policy choices of governments if and only if* $\bar{V}_{\tau}^{\mathcal{F}} = 0 = \bar{V}_{\tau^*}^{\mathcal{F}*}$ *when evaluated at politically optimal policies. Moreover, if* $\bar{V}_{\tau}^{\mathcal{F}} > 0$ *and/or* $\bar{V}_{\tau^*}^{\mathcal{F}*} > 0$*, it is possible that an efficient trade agreement must lead to reduced levels of market access for the member governments.*

It is notable that, according to Proposition 2, when fairness considerations involving own-country distributive and procedural justice concerns are present, only the presence of moral sentiments over own-country *tariff policy* will alter the purpose of a trade agreement: moral sentiments over own-country standards have no such impact. This point can be understood by noting that, in the complete absence of moral sentiments over own-country policies, it is tariffs, not standards, that are the first-best instrument for targeting import volumes, and as Bagwell and Staiger (2001) explain, this property plays a key role in shaping the problem that a trade agreement must solve in that setting. If citizens have moral sentiments only over own-country standards, this property continues to hold; but it no longer holds if citizens have moral sentiments over own-country tariff policy.

More broadly, our results for Case I confirm that the purpose and design of a trade agreement are not necessarily altered by the presence of concerns for fairness, and by implication these results suggest that the broad design features of the GATT/WTO may be well-suited to allow member governments to address at least some of the growing trade-related fairness concerns of their citizens. In particular, as Propositions 1-2 report, in Case I shallow integration is always a feasible approach to attaining the international efficiency frontier, though whether other features of the GATT/WTO design can be rationalized hinges on whether stringent claims extend to the level of own-country tariffs.

4.2 Case II: Own-Country Distributive Justice and Procedural Justice Everywhere

When the representative agent in each country has own-country distributive justice concerns and procedural justice concerns everywhere corresponding to the moral sentiments in the top left and top right boxes plus the bottom right

This definition of the political optimum deviates from that adopted by Bagwell and Staiger (1999, 2001, 2002) where the focus is on the international (material) cost-shifting that occurs with movements in $\tilde{p}^w$, as embodied in the terms V_{p^w} and $V^*_{p^w}$, and we therefore choose not to use it. But while under this alternative definition the purpose of a trade agreement can be said to be the elimination of terms-of-trade manipulation from unilateral policy choices, in the end it simply offers a different perspective on the same point that we emphasize in the text: when citizens have stringent claims over their own-country tariff levels, the purpose of a trade agreement goes beyond the elimination of international cost-shifting motives to include as well terms-of-trade manipulation for its impact on aggrievement.

box of Figure 1, we can write the total welfare functions of the domestic and foreign citizen respectively as

$$\begin{aligned} W = & V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ & - \mathcal{A}(\hat{V}^{\mathcal{F}}(s, \tau, s^*, \tau^*), V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ & \equiv \hat{W}(V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}}(s, \tau, s^*, \tau^*)) \end{aligned} \quad (27)$$

$$\begin{aligned} W^* = & V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ & - \mathcal{A}^*(\hat{V}^{\mathcal{F}*}(s^*, \tau^*, s, \tau), V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ & \equiv \hat{W}^*(V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}*}(s^*, \tau^*, s, \tau)) \end{aligned} \quad (28)$$

with $\hat{W}$ and $\hat{W}^*$ increasing in their first arguments and weakly decreasing in their second arguments.

As before we assume that each government chooses its policy instruments with the goal of maximizing the material utility of its representative citizen minus a weight ($\gamma \geq 0$ for the domestic government, $\gamma^* \geq 0$ for the foreign government) times the aggrievement suffered by its representative citizen. Formally G and G^* are now specified as follows:

$$\begin{aligned} G = & V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ & - \gamma \times \mathcal{A}(\hat{V}^{\mathcal{F}}(s, \tau, s^*, \tau^*), V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ & \equiv \hat{G}(V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}}(s, \tau, s^*, \tau^*), \gamma) \end{aligned} \quad (29)$$

$$\begin{aligned} G^* = & V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ & - \gamma^* \times \mathcal{A}^*(\hat{V}^{\mathcal{F}*}(s^*, \tau^*, s, \tau), V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\ & \equiv \hat{G}^*(V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}*}(s^*, \tau^*, s, \tau), \gamma^*) \end{aligned} \quad (30)$$

with $\hat{G}$ and $\hat{G}^*$ increasing in their first arguments and weakly decreasing in their second arguments.

Internationally efficient policies and deep integration Proceeding as before, the solution to the following program characterizes the internationally efficient choices of the domestic and foreign policies for Case II:

$$\begin{aligned} \max_{s,\tau,s^*,\tau^*} \quad & \hat{G}(V(s, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}}(s, \tau, s^*, \tau^*), \gamma) \\ & s.t. \\ \hat{G}^*(V^*(s^*, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \hat{V}^{\mathcal{F}*}(s^*, \tau^*, s, \tau), \gamma^*) & \geq \hat{G}^{*E} \end{aligned} \quad (31)$$

where $\hat{G}^{*E} \equiv \hat{G}^*(V^*(s^{*E}, p^*(\tau^{*E}, \tilde{p}^{wE}), \tilde{p}^{wE}), \hat{V}^{\mathcal{F}*}(s^{*E}, \tau^{*E}, s^E, \tau^E), \gamma^*)$.

There are two possibilities to consider. One possibility is that, when evaluated at any internationally efficient policies, we have $V_{\tau^*}^{\mathcal{F}} = 0 = \hat{V}_{s^*}^{\mathcal{F}}$ and $\hat{V}_s^{\mathcal{F}*} = 0 = \hat{V}_\tau^{\mathcal{F}*}$ so that neither the domestic citizen nor the foreign citizen is aggrieved over its trading partner's policy choices when these choices reach the international efficiency frontier. Under this possibility our characterization of the efficiency frontier for Case I given in (20)-(22) applies without modification. Here we focus on the other possibility that at least one of these derivatives is non-zero, so that for at least one citizen there is aggrievement over at least one policy chosen by the trading partner when these policy choices reach the international efficiency frontier. We will refer to this second possibility as the case where fairness considerations involving own-country distributive concerns and procedural justice concerns everywhere are both present *and* "relevant on the international efficiency frontier."

To characterize the international efficiency frontier under this second possibility, we will without loss of generality assume in what follows that $\hat{V}_{\tau^*}^{\mathcal{F}} \neq 0 \neq \hat{V}_{s^*}^{\mathcal{F}}$ and $\hat{V}_s^{\mathcal{F}*} \neq 0 \neq \hat{V}_\tau^{\mathcal{F}*}$ when evaluated at internationally efficient policies. As before, forming the Lagrangian associated with (31) and manipulating the four first order conditions with respect to s, τ, s^* and τ^* to eliminate the Lagrange multiplier yields the three conditions that define the international efficiency frontier for Case II:

$$\left[\frac{\tilde{p}^w \hat{G}_V V_p + \hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_\tau^{\mathcal{F}} + \left[\hat{G}_V V_s + \hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_s^{\mathcal{F}}\right] \times \frac{ds}{d\tau}|_{d\tilde{p}^w=0}}{\hat{G}_V \left[(\tau V_p + V_{p^w}) \frac{\partial \tilde{p}^w}{\partial \tau} + \tilde{p}^w V_p\right] + \hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_\tau^{\mathcal{F}}}\right] =$$

$$\left[\frac{\hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_\tau^{\mathcal{F}*}}{\hat{G}^*_{\hat{V}^*} \left[\frac{1}{\tau^*} V^*_{p^*} + V^*_{p^w}\right] \frac{\partial \tilde{p}^w}{\partial s}}\right] \times \left[\frac{d\tau}{ds}|_{d\hat{V}^{\mathcal{F}*}=0} - \frac{d\tau}{ds}|_{d\hat{G}=0}\right] \quad (32)$$

$$\left[\frac{\left(-\frac{p^*}{\tau^*}\right) \hat{G}^*_{V^*} V^*_{p^*} + \hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_{\tau^*}^{\mathcal{F}*} + \left[\hat{G}^*_{V^*} V^*_{s^*} + \hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_{s^*}^{\mathcal{F}*}\right] \times \frac{ds^*}{d\tau^*}|_{d\tilde{p}^w=0}}{\hat{G}^*_{V^*} \left[\left(\frac{1}{\tau^*} V^*_{p^*} + V^*_{p^w}\right) \frac{\partial \tilde{p}^w}{\partial \tau^*} + \left(-\frac{p^*}{\tau^*}\right) V^*_{p^*}\right] + \hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_{\tau^*}^{\mathcal{F}*}}\right] =$$

$$\left[\frac{\hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_{\tau^*}^{\mathcal{F}}}{\hat{G}_{\hat{V}} \left[\tau V_p + V_{p^w}\right] \frac{\partial \tilde{p}^w}{\partial s^*}}\right] \times \left[\frac{d\tau^*}{ds^*}|_{d\hat{V}^{\mathcal{F}}=0} - \frac{d\tau^*}{ds^*}|_{d\hat{G}^*=0}\right] \quad (33)$$

$$\left[\frac{\hat{G}_V \left[V_p \frac{dp}{d\tau} + V_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau}\right] + \hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_\tau^{\mathcal{F}}}{\hat{G}^*_{V^*} \left[\frac{1}{\tau^*} V^*_{p^*} + V^*_{p^w}\right] \frac{\partial \tilde{p}^w}{\partial \tau} + \hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_\tau^{\mathcal{F}*}}\right] =$$

$$\left[\frac{\hat{G}_V \left[\tau V_p + V_{p^w}\right] \frac{\partial \tilde{p}^w}{\partial \tau^*} + \hat{G}_{\hat{V}^{\mathcal{F}}} \hat{V}_{\tau^*}^{\mathcal{F}}}{\hat{G}^*_{V^*} \left[V^*_{p^*} \frac{dp^*}{d\tau^*} + V^*_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau^*}\right] + \hat{G}^*_{\hat{V}^{\mathcal{F}*}} \hat{V}_{\tau^*}^{\mathcal{F}*}}\right] \quad (34)$$

where in writing (32) and (33) we have used the fact that

$$\frac{d\tau}{ds}|_{d\hat{V}^{\mathcal{F}*}=0} = -\frac{\hat{V}_s^{\mathcal{F}*}}{\hat{V}_\tau^{\mathcal{F}*}} \text{ and } \frac{d\tau}{ds}|_{d\hat{G}=0} = -\frac{\hat{G}_V\left[(\tau V_p + V_{p^w})\frac{\partial \tilde{p}^w}{\partial s} + V_s\right] + \hat{G}_{\hat{V}^{\mathcal{F}}}\hat{V}_s^{\mathcal{F}}}{\hat{G}_V\left[(\tau V_p + V_{p^w})\frac{\partial \tilde{p}^w}{\partial \tau} + \tilde{p}^w V_p\right] + \hat{G}_{\hat{V}^{\mathcal{F}}}\hat{V}_\tau^{\mathcal{F}}}, \text{ and}$$

$$\frac{d\tau^*}{ds^*}|_{d\hat{V}^{\mathcal{F}}=0} = -\frac{\hat{V}_{s^*}^{\mathcal{F}}}{\hat{V}_{\tau^*}^{\mathcal{F}}} \text{ and } \frac{d\tau^*}{ds^*}|_{d\hat{G}^*=0} = -\frac{\hat{G}^*_{V^*}\left[(\frac{1}{\tau^*}V^*_{p^*} + V^*_{p^w})\frac{\partial \tilde{p}^w}{\partial s^*} + V^*_{s^*}\right] + \hat{G}^*_{\hat{V}^{\mathcal{F}*}}\hat{V}_{s^*}^{\mathcal{F}*}}{\hat{G}^*_{V^*}\left[(\frac{1}{\tau^*}V^*_{p^*} + V^*_{p^w})\frac{\partial \tilde{p}^w}{\partial \tau^*} + \left(-\frac{p^*}{\tau^*}\right)V^*_{p^*}\right] + \hat{G}^*_{\hat{V}^{\mathcal{F}*}}\hat{V}_{\tau^*}^{\mathcal{F}*}}.$$

Armed with the characterization of the international efficiency frontier in (32)-(34), we can now focus on the novel features that arise in this setting.

To this end, notice that the numerator of the term on the left-hand side of (32) gives the impact on the domestic government of a small increase in τ combined with a small change in s that holds the equilibrium relative world price $\tilde{p}^w$ fixed, the same interpretation as the expression on the left-hand side of (20) under Case I. And as the right-hand side of (32) indicates, if the foreign stringent-claims function $\hat{V}^{\mathcal{F}*}$ happened to satisfy $\frac{d\tau}{ds}|_{d\hat{V}^{\mathcal{F}*}=0} = \frac{d\tau}{ds}|_{d\hat{G}=0}$ at efficient policies, so that small changes in τ and s that left the domestic government indifferent would leave foreign stringent claims unchanged, then the right-hand side of (32) would be equal to zero, and (32) could be interpreted as the domestic "national" efficiency condition, just as we interpreted (20). But in general there is no reason to expect that $\hat{V}^{\mathcal{F}*}$ would satisfy this property. And this means that isolating a condition for international efficiency that only involves tradeoffs as perceived by the domestic government is no longer possible. An analogous observation from the perspective of the foreign government holds for (33) with regard to the domestic stringent claims function $\hat{V}^{\mathcal{F}}$. And this means that the logic of shallow integration that held in Case I no longer applies in Case II.

In particular, holding each government to adjustments in its mix of policies that do not alter the level of market access that it committed to in the tariff negotiations and hence do not alter the market clearing world price $\tilde{p}^w$ will *not* in general lead to choices that satisfy the efficiency conditions in (32) and (33). International efficiency in this case then requires that governments negotiate directly over their standards as well as their tariffs. We summarize with

Proposition 3 *When fairness considerations involving own-country distributive concerns and procedural justice concerns everywhere are present and relevant on the international efficiency frontier, a GATT-like shallow approach to efficient integration is infeasible. Instead, to achieve the international efficiency frontier, countries must negotiate directly over the policies of their trading partners that enter directly in the stringent claims of their citizens.*

As a comparison of Propositions 1 and 3 confirms, when fairness considerations spill over into procedural justice concerns in other countries, only a deep approach to integration – where governments negotiate directly over the policies that have triggered these procedural justice concerns – will suffice to allow countries to reach the international efficiency frontier.

It is illuminating to pause briefly and consider in more detail what these propositions imply for the interpretation of the design of a trade agreement such as the GATT/WTO. In particular, we can focus on the possibility that a surge in foreign exports leads to injury in the domestic market, and consider the design of the WTO Safeguards Agreement (the SG Agreement) as compared to the design of the WTO Agreement on Subsidies and Countervailing Measures (the SCM Agreement), either of which could become relevant in this circumstance.

We can consider in turn each of the two cases identified above, and ask whether the relevant agreement in the WTO is shallow or deep. Under Case I, the surge in foreign exports does not lead domestic citizens to become aggrieved on account of specific foreign policies that underlie that surge. This might describe a circumstance, for example, where the surge in foreign exports was due to a foreign technological improvement, or where a weather event in the foreign country was favorable to foreign exports.[28] In a circumstance such as this, Article XIX of GATT and the SG Agreement that reinforces it provides the relevant WTO rules. And consistent with our results for Case I, the SG Agreement is a shallow-integration agreement, in the sense that its purpose is simply to reign in the disciplines on a country's allowable tariff actions in response to a surge in foreign exports, as the second paragraph of the SG Agreement explains:

> Recognizing the need to clarify and reinforce the disciplines of GATT 1994, and specifically those of its Article XIX (Emergency Action on Imports of Particular Products), to re-establish multilateral control over safeguards and eliminate measures that escape such control;...

In particular, the SG Agreement is not an attempt to negotiate directly over behind-the-border policies; and as long as there are no fairness concerns associated with a specific foreign policy that gave rise to the foreign export surge considered here, our findings indicate that there is indeed no need for deep integration to handle this circumstance.

Now suppose instead that the surge in foreign exports was caused by a foreign export subsidy ($t^* < 0$), and that as in our Case II this leads domestic citizens to become aggrieved on account of the foreign export subsidy that underlies that surge. Our results for Case II indicate that direct negotiation over the foreign export subsidy (deep integration) is then required to reach the international efficiency frontier. And consistent with those results and in contrast to the SG Agreement, the SCM Agreement is a deep-integration Agreement, as evidenced by the fact that the agreement begins with an elaborate attempt to define a subsidy in order to then place disciplines on it (footnotes omitted):[29]

[28] Recall that in defining the equilibrium relative world price $\tilde{p}^w(s, \tau, s^*, \tau^*)$ we have held technologies, endowments and preferences fixed; a weather shock would the shift the $\tilde{p}^w$ function, which is what we are alluding to in the text.

[29] The SCM Agreement covers both export subsidies and subsidies to domestic production. As we have noted, the former are captured in our model by a negative import tariff; to capture production subsidies would require additional notation along the lines of Bagwell and Staiger (2006) who allow for separate producer and consumer prices within each country. But it is

Article 1: Definition of a Subsidy

1.1 For the purpose of this Agreement, a subsidy shall be deemed to exist if:

(a)(1) there is a financial contribution by a government or any public body within the territory of a Member (referred to in this Agreement as "government"), i.e. where:

(i) a government practice involves a direct transfer of funds (e.g. grants, loans, and equity infusion), potential direct transfers of funds or liabilities (e.g. loan guarantees);

(ii) government revenue that is otherwise due is foregone or not collected (e.g. fiscal incentives such as tax credits);

(iii) a government provides goods or services other than general infrastructure, or purchases goods;

(iv) a government makes payments to a funding mechanism, or entrusts or directs a private body to carry out one or more of the type of functions illustrated in (i) to (iii) above which would normally be vested in the government and the practice, in no real sense, differs from practices normally followed by governments; or

(a)(2) there is any form of income or price support in the sense of Article XVI of GATT 1994; and

(b) a benefit is thereby conferred.

Hence, under the assumption that injury due to trade flows is deemed unfair when those trade flows arise due to government subsidies but not when they arise due to "non-governmental" market forces, the shallow/deep differences across the SG Agreement and the SCM Agreement can be interpreted through the lens of our modeling framework as reflecting the relevance of fairness considerations in the design of the WTO.[30]

5 Heterogeneous Agents

We now extend our results to a setting with heterogeneous agents. This requires that we first express government objectives in light of the underlying citizen total welfare functions given by (11)-(14). We will assume that each government maximizes a weighted sum of the welfares of its citizens, where these weights are positive but otherwise unrestricted and are therefore consistent with a wide set of distributional/political economy motivations as in Bagwell and Staiger (1999, 2001), and where as before we capture the sensitivity of the domestic and foreign

straightforward to show that the results we derive here extend without qualification to such a model.

[30] One might think that the WTO Anti-Dumping Agreement (AD Agreement) should also be considered in this comparison, along side the SCM Agreement. But the AD Agreement focuses on disciplining the use of anti-dumping duties only; it does not attempt to impose restrictions on the act of dumping itself. This is because dumping refers to an action taken by a firm, not a government, and the WTO is a government-to-government agreement whose rules are restricted to disciplining the policy actions that governments take.

governments to the fairness concerns of their citizens with the parameters γ and γ^* respectively. We also assume that the transfers that each government offers to its citizens satisfy an adding up constraint, so that $\sum_{i=1}^{N} T^i = 0$ and $\sum_{i=1}^{N^*} T^{i*} = 0$ where N and N^* denote the number of citizens in the domestic and foreign country, respectively.

To write down the domestic government's objective function, we first partition domestic citizens by their stringent claims functions, creating the following three sets that mirror the results of Propositions 1-3:

$$\begin{aligned}
\Omega_{Ia} &\equiv \{i|\ V^{\mathcal{F}i} = \ddot{V}^{\mathcal{F}i}(s, T^i, p(\tau, \tilde{p}^w))\} \\
\Omega_{Ib} &\equiv \{i|\ V^{\mathcal{F}i} = \bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w))\} \\
\Omega_{II} &\equiv \{i|\ V^{\mathcal{F}i} = \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*)\}.
\end{aligned}$$

As defined, Ω_{Ia} is the set of domestic citizens whose stringent claims take the form $V^{\mathcal{F}i} = \ddot{V}^{\mathcal{F}i}(s, T^i, p(\tau, \tilde{p}^w))$, and hence the set of citizens whose fairness concerns according to Propositions 1 and 2 would leave the purpose and design of a trade agreement unaffected; Ω_{Ib} is the set of domestic citizens whose stringent claims take the form $V^{\mathcal{F}i} = \bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w))$, and hence the set of citizens whose fairness concerns according to Propositions 1 and 2 would alter the purpose of a trade agreement if $\bar{V}_{\tau}^{\mathcal{F}i} \neq 0$ and/or $\bar{V}_{\tau^*}^{\mathcal{F}*i} \neq 0$ when evaluated a politically optimal policies but leave the viability of shallow integration intact. Together the sets Ω_{Ia} and Ω_{Ib} span the Case I results reported in Propositions 1 and 2. And Ω_{II} is the set of domestic citizens whose stringent claims take the form $V^{\mathcal{F}i} = \hat{V}^{\mathcal{F}i}(s, T^i, \tau, s^*, \tau^*)$, and hence the set of citizens whose fairness concerns according to Proposition 3 would alter the purpose of a trade agreement *and* interfere with the viability of shallow integration if these concerns are relevant on the international efficiency frontier, corresponding to Case II.

Using these sets, we can then express the domestic government objective function for the heterogeneous agent setting as

$$\begin{aligned}
\Gamma = &\sum_{i=1}^{N} \alpha^i V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\
&- \gamma \sum_{i \in \Omega_{Ia}} \alpha^i \mathcal{A}^i(\ddot{V}^{\mathcal{F}i}(s, T^i, p(\tau, \tilde{p}^w)), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\
&- \gamma \sum_{i \in \Omega_{Ib}} \alpha^i \mathcal{A}^i(\bar{V}^{\mathcal{F}i}(s, T^i, \tau, p(\tau, \tilde{p}^w)), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))) \\
&- \gamma \sum_{i \in \Omega_{II}} \alpha^i \mathcal{A}^i(\hat{V}^{\mathcal{F}i}(s, \tau, T^i, s^*, \tau^*), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))
\end{aligned} \tag{35}$$

where α^i is the weight that the domestic government places on the welfare of

citizen i. Defining the analogous sets for the foreign country

$$\begin{aligned}\Omega^*_{Ia} &\equiv \{i|\ V^{\mathcal{F}*i} = \ddot{V}^{\mathcal{F}*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w))\}\\ \Omega^*_{Ib} &\equiv \{i|\ V^{\mathcal{F}*i} = \bar{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, p^*(\tau^*, \tilde{p}^w))\}\\ \Omega^*_{II} &\equiv \{i|\ V^{\mathcal{F}*i} = \hat{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, s, \tau)\},\end{aligned}$$

we can then express the foreign government objective function for the heterogeneous agent setting as

$$\begin{aligned}\Gamma^* = &\sum_{i=1}^{N^*} \alpha^{*i} V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))\\ -\gamma^* &\sum_{i\in\Omega^*_{Ia}} \alpha^{*i} \mathcal{A}^i(\ddot{V}^{\mathcal{F}*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w)), V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))\\ -\gamma^* &\sum_{i\in\Omega^*_{Ib}} \alpha^{*i} \mathcal{A}^{*i}(\bar{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, p^*(\tau^*, \tilde{p}^w)), V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))\\ -\gamma^* &\sum_{i\in\Omega^*_{II}} \alpha^{*i} \mathcal{A}^{*i}(\hat{V}^{\mathcal{F}*i}(s^*, T^{*i}, \tau^*, s, \tau), V^{*i}(s^*, T^{*i}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))\end{aligned} \tag{36}$$

where α^{*i} is the weight that the foreign government places on the welfare of citizen i.

With the expressions for domestic and foreign government objectives in (35) and (36), arguments analogous to those that led to Propositions 1-3 lead to the following:

Proposition 4 *In a heterogeneous agent setting, a GATT-like shallow approach to efficient integration remains feasible in the presence of trade-related fairness concerns if and only if the sets Ω_{II} and Ω^*_{II} are empty, or if they are non-empty but $\hat{V}^{\mathcal{F}i}_{\tau^*} = 0 = \hat{V}^{\mathcal{F}i}_{s^*}$ and $\hat{V}^{\mathcal{F}*i}_{s} = 0 = \hat{V}^{\mathcal{F}*i}_{\tau}$ for i in these sets when evaluated at any internationally efficient policies. If the sets Ω_{Ib} and Ω^*_{Ib} are also empty, then the purpose of a trade agreement is unchanged by the presence of trade-related fairness concerns, while if the sets Ω_{Ib} and Ω^*_{Ib} are non-empty then the purpose of a trade agreement is changed if and only if for some citizen i in these sets $\bar{V}^{\mathcal{F}i}_{\tau} \neq 0$ and/or $\bar{V}^{\mathcal{F}*i}_{\tau^*} \neq 0$ when evaluated at the politically optimal policies, and in this case fairness concerns may lead an efficient trade agreement to restrict trade below noncooperative levels.*

We can also state the following:

Corollary *The purpose and design of a trade agreement will be altered by fairness considerations if and only if some citizens somewhere would be aggrieved on the international efficiency frontier by treatment that they deem to be unfair on procedural justice grounds with respect to own tariffs or policies of their trading partners.*

An interesting feature of the Corollary to Proposition 4 is the decisive role that aggrievement *on the international efficiency frontier* plays in determining the purpose and design of a trade agreement. This raises the possibility that different points on the efficiency frontier could imply distinct requirements for the trade agreement that would be needed to reach them. Suppose, for example, that in the Nash equilibrium that would arise in the absence of an international policy agreement, poor countries are choosing weak labor standards that lead the citizens of rich countries to feel aggrieved. A point on the efficiency frontier where poor countries are given little of the surplus from international agreement is unlikely to involve poor-country choices of labor standards that would rise to the point of eliminating the aggrievement of rich country citizens; and by the Corollary to Proposition 4, reaching this point would not be possible with a GATT-like shallow approach to integration. But a point on the efficiency frontier where poor countries are given a greater share of the surplus from international agreement *might* be consistent with shallow integration, if the added income received by poor countries as a result of the more favorable terms of this agreement led them to choose standards that citizens in rich countries found to be morally acceptable and no longer a cause for aggrievement.[31]

Figure 2 summarizes the taxonomy of the results reported in Proposition 4, described from the perspective of the domestic country but understood to apply to both countries, and assessed from a position on the international efficiency frontier. A GATT-like shallow approach to integration remains feasible if the moral sentiments of all citizens fall within the top left and right boxes of the figure, and the purpose of a trade agreement is unchanged by fairness considerations as well as long as the tariff level itself is not a moral consideration. For these cases – signified in Figure 2 by the marker "FTA≡TA" – fairness concerns have no bearing on the purpose or design of a trade agreement. When the tariff level itself is a moral concern, the purpose of a trade agreement changes, which in Figure 2 is signified by the marker "FTA≠TA," but as long as moral sentiments are confined to the top left and right boxes of Figure 2 shallow integration is still feasible. Finally, when moral sentiments are not confined to the top two boxes of Figure 2, but also fall in the bottom right box, these moral sentiments both alter the purpose of a trade agreement and interfere with the ability of a shallow approach to integration to achieve the international efficiency frontier (we will comment on the bottom left box of Figure 2 in section 6).

[31] More broadly, in light of the Corollary to Proposition 4, it is conceivable that efficient shallow integration could lead to greater (vertical) foreign direct investment from rich countries into poor countries, and that the higher standards of rich countries might be spread to poor countries in the process, thereby alleviating the fairness concerns of rich country citizens over low standards in poor countries. Or it could be that shallow integration combined with imperfect property rights over negotiated market access has lead to a regulatory race to the bottom under shallow integration as described in Bagwell and Staiger (2001), and that the standards in poor countries chosen as a result of this race to the bottom are the cause of aggrievement in rich countries. In this case, it is conceivable that more perfect shallow integration (i.e., shallow integration with more effective rules to prevent the erosion of market access commitments with behind-the-border policies) could alleviate the race to the bottom and thereby address fairness concerns over standards.

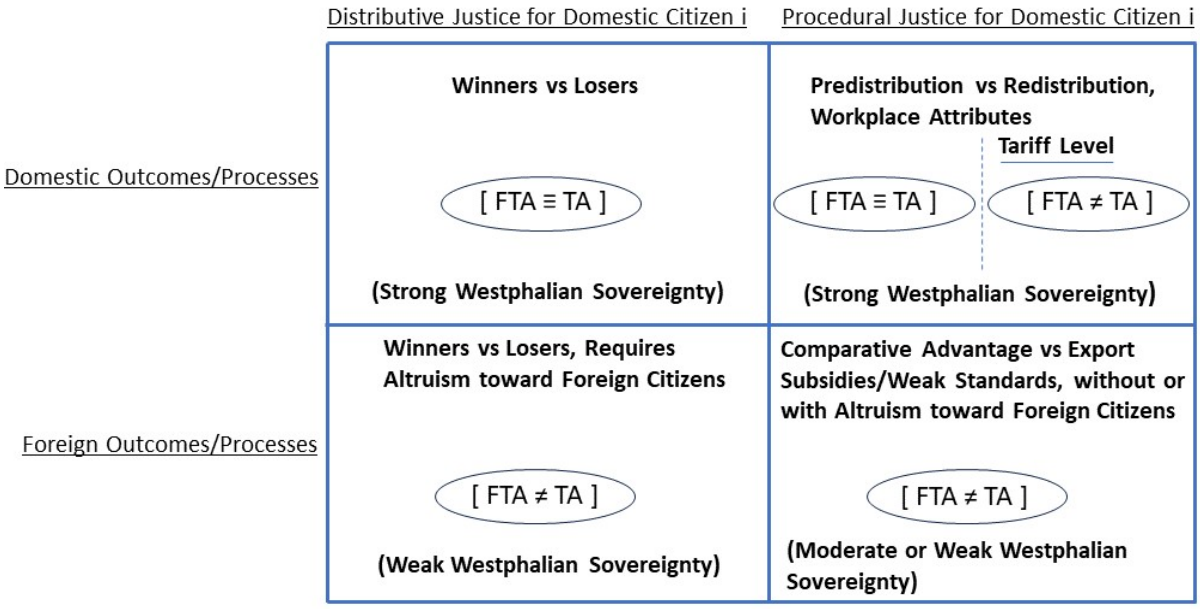

Figure 2: The Implications of Moral Sentiments for Fair Trade Agreements

We can also relate the taxonomy described in Figure 2 to recent work by international legal scholars. In addition to Meyer (forthcoming) who draws the distinction between production and consumption jurisdiction approaches to international trade law (see notes 19 and 20), Shaffer (2024) considers a related notion of "flanking policies" which was first described by Pauwelyn and Sieber-Gasser (2024) and which refers to additional policy interventions that may be needed in trade agreements as a result of (among other things) moral sentiments. As defined by Pauwelyn and Sieber-Gasser, flanking policies are "policies that can mitigate negative effects of trade liberalization, or the concerns of domestic stakeholders regarding said effects, or both, and that are legally or factually linked to such trade liberalization." As these authors note, flanking policies could either be inherently a national affair – "first generation" flanking policies – or a transnational matter – "second generation" flanking policies – corresponding loosely to the policies that would be needed in response to the moral sentiments in our Case I and Case II respectively.

Accordingly, in the terminology of Pauwelyn and Sieber-Gasser (2024), first generation flanking policies can be implemented unilaterally by each government's choice of its own domestic policies and are therefore consistent with shallow integration (our Case I, top left and right boxes of Figure 2), while second generation flanking policies must be negotiated across governments and hence require deep integration over domestic policies to implement (our Case II, bottom right box of Figure 2). And the initial acceptance of first generation

flanking policies as sufficient to accompany trade agreements that has increasingly given way to the view that second generation flanking policies must also accompany trade agreements reflects, in the terminology of Meyer, an evolution from a production jurisdiction norm to a consumption jurisdiction norm.

Finally, it is worth emphasizing from a practical perspective the range and nature of the fairness concerns whose presence would *not* change the purpose or design of a trade agreement according to our findings, and hence in whose presence the broad design features of the GATT/WTO would continue to be supported by economic arguments such as those in Bagwell and Staiger (1999, 2001, 2002), and also those fairness concerns that *would* overturn these arguments. While not translating the language of our model into common parlance, we can nevertheless gain an appreciation of these issues by considering the answers to the following five hypothetical questions posed to (domestic) citizen i about the nature of his stringent claims, that is, about what he feels he is owed:

> When it comes to being treated fairly in a world with foreign trade pressures:
>
> i) Do you feel you are owed a level of *material utility* that is independent of policy and context and that foreign trade pressures could either create or threaten?
>
> ii) Does the level of material utility that you feel you are owed depend on the level of *domestic standards* or *domestic material outcomes other than utility*?
>
> iii) Does the level of material utility that you feel you are owed depend on the level of *domestic tariffs*?
>
> iv) Does the level of material utility that you feel you are owed depend on *the foreign government's overall policy stance* and how that policy stance contributes to the level of foreign trade pressure?
>
> v) Holding fixed the level of foreign trade pressure, does the level of material utility that you feel you are owed depend on the *specific mix of foreign policies* that lie behind this foreign trade pressure?

As long as the answer to the last question listed above is "No," an affirmative answer to any or all of the first four questions is consistent with the stringent claims function in (8), and hence Case I applies. In this case shallow integration is always a feasible approach to attaining the international efficiency frontier, though whether other features of the GATT/WTO design can be rationalized or not hinges on the answer to the third question above. It is only when the fifth question is answered in the affirmative that Case II becomes relevant, and only in that case that fairness concerns would have implications for the purpose and design of a trade agreement that are fundamentally inconsistent with the broad design features of the GATT/WTO (and even then only if the fairness concerns described in the fifth question are relevant on the international efficiency frontier).

5.1 Concern for personal dignity

With Proposition 4 providing a generalization of our representative agent results to the heterogeneous agent world, we now turn to focus on the new element introduced by the heterogeneous agent model, namely, the role of transfers and the possibility that the citizens of a country might have a preference for predistribution over redistribution out of concerns for personal dignity. Concerns for personal dignity that might give rise to such preferences have been emphasized recently in both academic studies and the popular press as potentially relevant for understanding the backlash against globalization and the political realignments over trade policy that have taken place in the United States and elsewhere (see, for example, Kuziemko, Longuet-Marx and Naidu, 2023, and Porter, 2024). Here we explore how concerns for dignity play out – in terms of noncooperative policy choices, efficient policy choices, and the purpose of a fair trade agreement – relative to what would be expected in the absence of dignity concerns.

To focus on these novel issues, we abstract from the other procedural justice concerns highlighted in our representative agent model and assume that, in addition to distributive justice concerns, there is only one possible procedural justice concern: domestic and foreign citizens may (or may not) have dignity concerns about receiving transfers. That is, we restrict attention to stringent claims that take the form of $\dddot{V}^{\mathcal{F}i}(T^i)$ for domestic citizens and $\dddot{V}^{\mathcal{F}*i}(T^{*i})$ for foreign citizens. In terms of our taxonomy above, this puts us squarely in the case where only the sets Ω_{Ia} and Ω^*_{Ia} are non-empty, and therefore by Proposition 4 we know that the purpose of a trade agreement is unchanged by the presence of concerns for personal dignity. Nevertheless, it is informative to see how these moral sentiments impact the workings of the model.

Consider first how the noncooperative tariffs and standards are impacted by the presence of dignity concerns. For purposes of illustration, we focus on the domestic government's policy choices. The best-response noncooperative policy choices of the domestic government given any foreign policies s^* and τ^* are the solutions to the following program:

$$\max_{s,\tau,\{T^i\}} \sum_{i=1}^{N} \alpha^i V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) - \gamma \sum_{i \in \Omega_{Ia}} \alpha^i \mathcal{A}^i(\dddot{V}^{\mathcal{F}i}(T^i), V^i(s, T^i, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))$$

$$s.t. \quad \sum_{i=1}^{N} T^i \leq 0. \tag{37}$$

In the case where there are no dignity concerns among domestic citizens associated with transfers so that $\dddot{V}^{\mathcal{F}i}_{T^i} \equiv 0$ for $i \in \Omega_{Ia}$, the first-order conditions

associated with the domestic government's choice of transfers are given by

$$\begin{aligned} \alpha^i \left(1-\gamma \mathcal{A}^i_{V^i}\right) &= \lambda/V^i_{T^i} \quad \text{for } i \in \Omega_{Ia} \\ \alpha^i &= \lambda/V^i_{T^i} \quad \text{for } i \notin \Omega_{Ia} \end{aligned} \tag{38}$$

where λ is the Lagrange multiplier on the constraint in (37). The first-order conditions for the tariff and the standard are given respectively by

$$\sum_{i \notin \Omega_{Ia}} \alpha^i \left[V^i_p \frac{dp}{d\tau} + V^i_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau}\right] + \sum_{i \in \Omega_{Ia}} \alpha^i \left(1-\gamma \mathcal{A}^i_{V^i}\right) \left[V^i_p \frac{dp}{d\tau} + V^i_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau}\right] = 0 \tag{39}$$

$$\sum_{i \notin \Omega_{Ia}} \alpha^i \left[V^i_s + \left(\tau V^i_p + V^i_{p^w}\right) \frac{\partial \tilde{p}^w}{\partial s}\right] + \sum_{i \in \Omega_{Ia}} \alpha^i \left(1-\gamma \mathcal{A}^i_{V^i}\right) \left[V^i_s + \left(\tau V^i_p + V^i_{p^w}\right) \frac{\partial \tilde{p}^w}{\partial s}\right] = 0.$$

But substituting (38) into (39) and simplifying yields

$$\begin{aligned} \sum_{i=1}^{N} \frac{\left[V^i_p \frac{dp}{d\tau} + V^i_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau}\right]}{V^i_{T^i}} &= 0 \\ \sum_{i=1}^{N} \frac{\left[V^i_s + \left(\tau V^i_p + V^i_{p^w}\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{V^i_{T^i}} &= 0. \end{aligned} \tag{40}$$

Using (5), Roy's identity, the efficiency of competitive production and the fact that

$$I^i(s, p(\tau, \tilde{p}^w), \tilde{p}^w) \equiv \theta^i \times I(s, p(\tau, \tilde{p}^w), \tilde{p}^w)$$

where θ^i denotes citizen i's share of domestic national income with $\sum_{i=1}^{N} \theta^i \equiv 1$ and where $I(s, p(\tau, \tilde{p}^w), \tilde{p}^w)$ is the domestic national income measured in units of good y defined by

$$I(s, p(\tau, \tilde{p}^w), \tilde{p}^w) = pQ_x(s, p(\tau, \tilde{p}^w)) + Q_y(s, p(\tau, \tilde{p}^w)) + (p - \tilde{p}^w) \times M_x(s, p(\tau, \tilde{p}^w), \tilde{p}^w),$$

we can rewrite the top line of (40) as

$$\sum_{i=1}^{N} \frac{\left[V^i_p \frac{dp}{d\tau} + V^i_{p^w} \frac{\partial \tilde{p}^w}{\partial \tau}\right]}{V^i_{T^i}} =$$

$$\left[t\tilde{p}^w \times \frac{\partial M_x(s, p(\tau, \tilde{p}^w), \tilde{p}^w)}{\partial p}\right] \frac{dp}{d\tau} + \left[t\tilde{p}^w \frac{\partial M_x(s, p(\tau, \tilde{p}^w), \tilde{p}^w)}{\partial p^w} - M_x(s, p(\tau, \tilde{p}^w), \tilde{p}^w)\right] \frac{\partial \tilde{p}^w}{\partial \tau} = 0,$$

which can be manipulated to yield

$$t^R = \frac{1}{\frac{E^*_x}{\eta_{p^w}}},$$

where we use t^R to denote the best-response tariff of the domestic government and where $\eta_{p^w}^{E_x^*}$ is the elasticity of foreign export supply with respect to the relative world price p^w.

Therefore, when the domestic government has lump-sum transfers and there are no dignity concerns associated with transfers so that $\dddot{V}_{T^i}^{\mathcal{F}i} \equiv 0$ for $i \in \Omega_{Ia}$, the noncooperative tariff is simply the Johnson (1953-54) optimal tariff, with transfers then used by the domestic government to redistribute national income optimally in light of the citizen-level weights α^i in the domestic government's objective function. And using similar arguments, it can be shown that the noncooperative standard is then set efficiently given the trade volume implied by t^R, in line with our results above.

Now consider the case where there are dignity concerns among domestic citizens associated with transfers, so that $\dddot{V}_{T^i}^{\mathcal{F}i} > 0$ for $i \in \Omega_{Ia}$. In this case the first-order conditions associated with the domestic government's choice of transfers become

$$\begin{aligned} \alpha^i \left(1 - \gamma \mathcal{A}_{V^i}^i\right) &= \left[\lambda + \alpha^i \gamma \mathcal{A}_{\dddot{V}^{\mathcal{F}i}}^i \dddot{V}_{T^i}^{\mathcal{F}i}\right] / V_{T^i}^i \quad \text{for } i \in \Omega_{Ia} \qquad (41) \\ \alpha^i &= \lambda / V_{T^i}^i \quad \text{for } i \notin \Omega_{Ia}, \end{aligned}$$

while the first-order conditions associated with the domestic government's choice of tariff and standard are unchanged and given by (39). Substituting (41) into (39) and simplifying yields

$$\begin{aligned} \sum_{i=1}^{N} \frac{\left[V_p^i \frac{dp}{d\tau} + V_{p^w}^i \frac{\partial \tilde{p}^w}{\partial \tau}\right]}{V_{T^i}^i} &= -\sum_{i \in \Omega_{Ia}} \left[\alpha^i \gamma \mathcal{A}_{\dddot{V}^{\mathcal{F}i}}^i \dddot{V}_{T^i}^{\mathcal{F}i}\right] \frac{\left[V_p^i \frac{dp}{d\tau} + V_{p^w}^i \frac{\partial \tilde{p}^w}{\partial \tau}\right]}{\lambda V_{T^i}^i} \qquad (42) \\ \sum_{i=1}^{N} \frac{\left[V_s^i + \left(\tau V_p^i + V_{p^w}^i\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{V_{T^i}^i} &= -\sum_{i \in \Omega_{Ia}} \left[\alpha^i \gamma \mathcal{A}_{\dddot{V}^{\mathcal{F}i}}^i \dddot{V}_{T^i}^{\mathcal{F}i}\right] \frac{\left[V_s^i + \left(\tau V_p^i + V_{p^w}^i\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{\lambda V_{T^i}^i}. \end{aligned}$$

With $\dddot{V}_{T^i}^{\mathcal{F}i} > 0$ for $i \in \Omega_{Ia}$, the right-hand side of the top line in (40) will be strictly negative if an increase in the domestic tariff increases the real incomes of domestic citizens $i \in \Omega_{Ia}$, and it will be strictly positive if an increase in the domestic tariff decreases the real incomes of domestic citizens $i \in \Omega_{Ia}$.

Hence, comparing (40) to (42), it is apparent that when $\dddot{V}_{T^i}^{\mathcal{F}i} > 0$ for $i \in \Omega_{Ia}$ the domestic government will move its noncooperative tariff away from $t^R = 1/\eta_{p^w}^{E_x^*}$ in the direction that raises the real incomes of domestic citizens $i \in \Omega_{Ia}$. This reflects the fact that with $\dddot{V}_{T^i}^{\mathcal{F}i} > 0$ and citizen i therefore feeling a loss of dignity when he receives a government transfer, the government can no longer use transfers as effectively for the purpose of responding to the aggrievement of these citizens and therefore begins to enlist its tariff for this purpose as well.

What about the nature of the difference between efficient and noncooperative policies and therefore the purpose of a fair trade agreement? Following analogous arguments to those we made above, it can be confirmed that whatever the

value of $\dddot{V}^{\mathcal{F}i}_{T^i}$ for $i \in \Omega_{Ia}$, in the setting considered here, the purpose of a trade agreement remains the same, namely, to eliminate international cost-shifting from the unilateral policy choices of governments. This implies in turn that politically optimal tariffs and standards, defined as the tariffs and standards that would be chosen unilaterally by each government if it did not value its ability to shift costs onto its trading partner through terms-of-trade manipulation, are efficient.[32]

For the domestic government, the politically optimal policies are defined by its best-response policies under the assumption that $V^i_{p^w} \equiv 0$ for all i. When the domestic government has lump-sum transfers and there are no dignity concerns associated with transfers so that $\dddot{V}^{\mathcal{F}i}_{T^i} \equiv 0$ for $i \in \Omega_{Ia}$, the domestic politically optimal tariff and standard are defined using (40) by the following:

$$\begin{aligned} \sum_{i=1}^{N} \frac{\left[V^i_p \frac{dp}{d\tau}\right]}{V^i_{T^i}} &= \left[t\tilde{p}^w \times \frac{\partial M_x(s, p(\tau, \tilde{p}^w), \tilde{p}^w)}{\partial p}\right] \frac{dp}{d\tau} = 0 \qquad (43) \\ \sum_{i=1}^{N} \frac{\left[V^i_s + \left(\tau V^i_p\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{V^i_{T^i}} &= 0. \end{aligned}$$

As the top line of (43) implies, the politically optimal tariff for the domestic government will be zero when its citizens lack dignity concerns over the reception of transfers. This reflects the fact that in this case the noncooperative tariff is motivated completely by terms-of-trade manipulation (i.e., it is the Johnson optimal tariff), and when that motivation is removed the entire tariff is removed.

On the other hand, when dignity concerns associated with transfers are present, so that $\dddot{V}^{\mathcal{F}i}_{T^i} > 0$ for $i \in \Omega_{Ia}$, the domestic politically optimal tariff and standard are defined using (42) by

$$\begin{aligned} \sum_{i=1}^{N} \frac{\left[V^i_p \frac{dp}{d\tau}\right]}{V^i_{T^i}} &= -\sum_{i \in \Omega_{Ia}} \left[\alpha^i \gamma \mathcal{A}^i_{\dddot{V}^{\mathcal{F}i}} \dddot{V}^{\mathcal{F}i}_{T^i}\right] \frac{\left[V^i_p \frac{dp}{d\tau}\right]}{\lambda V^i_{T^i}} \qquad (44) \\ \sum_{i=1}^{N} \frac{\left[V^i_s + \left(\tau V^i_p\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{V^i_{T^i}} &= -\sum_{i \in \Omega_{Ia}} \left[\alpha^i \gamma \mathcal{A}^i_{\dddot{V}^{\mathcal{F}i}} \dddot{V}^{\mathcal{F}i}_{T^i}\right] \frac{\left[V^i_s + \left(\tau V^i_p\right) \frac{\partial \tilde{p}^w}{\partial s}\right]}{\lambda V^i_{T^i}}. \end{aligned}$$

With $\dddot{V}^{\mathcal{F}i}_{T^i} > 0$ for $i \in \Omega_{Ia}$, the term on the right-hand side of the top line of (44) is strictly negative (positive) if an increase (decrease) in the tariff increases the real incomes of domestic citizens $i \in \Omega_{Ia}$. Hence, as a comparison of the top lines of (43) and (44) confirm, if an increase (decrease) in the tariff increases the real incomes of domestic citizens $i \in \Omega_{Ia}$, the politically optimal domestic tariff will be strictly positive (negative) when $\dddot{V}^{\mathcal{F}i}_{T^i} > 0$ for $i \in \Omega_{Ia}$ and dignity concerns associated with transfers are present. In effect, under politically optimal choices the terms-of-trade-manipulation component of the noncooperative

[32] Under our assumption of identical and homothetic preferences within each country, transfers have no terms-of-trade impacts and so the transfers implied by the noncooperative first-order conditions above are automatically consistent with international efficiency.

tariff is removed, and what is left is the "legitimate" part of the noncooperative tariff that is meant to address aggrievement.

We summarize with:

Proposition 5 *The presence of dignity concerns diminishes government reliance on transfers as a means to address aggrievement associated with unfair treatment and instead promotes reliance on tariffs for this purpose, both under noncooperative policy choices and in the context of a fair trade agreement. But these concerns by themselves do not change the purpose of a trade agreement, and hence these concerns by themselves do not have implications for the design of a fair trade agreement.*

5.2 The impact of a fair trade agreement on aggrievement

We next show that a fair trade agreement will not necessarily reduce aggrievement related to feelings of unfair treatment from trade. Since this is a possibility result, we can do this in the setting described just above for the case where dignity concerns are present, but with three additional assumptions that help to make the point in a simple way.

First, we abstract from standards and assume that governments choose only tariffs and transfers. For simplicity we will simply assume that the domestic standard is fixed exogenously at the level $\bar{s}$, while the foreign standard is fixed exogenously at the level $\bar{s}^*$. Second, we assume that the two countries are symmetric, in the sense that a move from noncooperative Nash tariffs to efficient politically optimal tariffs does not alter the equilibrium relative world price. And third, we assume that the citizens in the sets Ω_{Ia} and Ω^*_{Ia} are all citizens whose real factor income rises with the tariff (e.g., either these citizens derive their factor income from ownership of factors of production that are specific to the country's import-competing sector, or they are owners of the factor used intensively in that sector). In this setting, we will show that, if beginning from Nash tariffs domestic and foreign citizens in the respective sets Ω_{Ia} and Ω^*_{Ia} are aggrieved, then a fair trade agreement that implements efficient politically optimal tariffs must *increase* the level of aggrievement in each country provided that the stringent claims functions $\dddot{V}^{\mathcal{F}i}(T^i)$ and $\dddot{V}^{\mathcal{F}*i}(T^{*i})$ are sufficiently convex so that citizen i's loss of dignity rises at a sufficiently rapid rate with the size of the transfer he receives.

Given our symmetry assumption, we can focus on aggrievement in the domestic country, which under Nash policies is given by

$$\sum_{i\in\Omega_{Ia}} \alpha^i \mathcal{A}^i(\dddot{V}^{\mathcal{F}i}(T^{iN}), V^i(\bar{s}, T^{iN}, p(\tau^N, \tilde{p}^{wN}), \tilde{p}^{wN}))$$

where $\tilde{p}^{wN} \equiv \tilde{p}^w(\bar{s}, \tau^N, \bar{s}, \tau^{*N})$ and T^{iN}, τ^N and τ^{*N} are Nash policies. A trade agreement that implements the politically optimal tariffs implies that tariffs will be lowered from their Nash levels, and our symmetry assumption implies that the relative world price is unaffected by these tariff cuts so that $\tilde{p}^{wN} \equiv$

$\tilde{p}^w(\bar{s}, \tau^N, \bar{s}, \tau^{*N}) = \tilde{p}^w(\bar{s}, \tau^{PO}, \bar{s}, \tau^{*PO}) \equiv \tilde{p}^{wPO}$ where τ^{PO} and τ^{*PO} are politically optimal tariffs; this in turn means that $p(\tau^{PO}, \tilde{p}^{wPO}) < p(\tau^N, \tilde{p}^{wN})$, so factor incomes for domestic citizens $i \in \Omega_{Ia}$ will be lower as a result of the trade agreement. We will also assume for simplicity that tariff revenue collected under the politically optimal tariff is lower than that collected under Nash tariffs, an assumption that is guaranteed to hold if the elasticity of foreign export supply is sufficiently low.[33] This implies that, in the absence of a change in the transfer T^i, the material utility V^i of every domestic citizen $i \in \Omega_{Ia}$ would fall; and given that $\mathcal{A}^i$ is strictly decreasing in V^i when $\mathcal{A}^i$ is strictly positive as we have assumed under Nash tariffs, this would in turn imply an increase in aggrievement in the domestic country.

Of course, as its tariff is reduced under the trade agreement, the domestic government *will* alter its choice of T^i, and in fact the government will to some extent increase T^i for each $i \in \Omega_{Ia}$ in order to make up for the negative distributional consequences on these citizens of its lower tariff. But we now argue that it could never be optimal for the government to increase T^i by enough to keep $\mathcal{A}^i$ from rising, as long as $\dddot{V}^{\mathcal{F}i}(T^i)$ is sufficiently convex.

To facilitate the argument, it is helpful to rewrite slightly the domestic government's first-order conditions for transfers in (41) in the equivalent form

$$\begin{aligned} \alpha^i \left[\left(1 - \gamma \mathcal{A}^i_{V^i}\right) V^i_{T^i} - \gamma \mathcal{A}^i_{\dddot{V}^{\mathcal{F}i}} \dddot{V}^{\mathcal{F}i}_{T^i} \right] &= \lambda \text{ for } i \in \Omega_{Ia} \\ \alpha^i V^i_{T^i} &= \lambda \text{ for } i \notin \Omega_{Ia}. \end{aligned} \tag{45}$$

Suppose, then, that in response to reducing its tariff from τ^N to τ^{PO}, the domestic government increased T^i sufficiently for some $i \in \Omega_{Ia}$ so that $\mathcal{A}^i$ remained fixed. By the definition of $\mathcal{A}^i$, this then implies that $\mathcal{A}^i_{V^i}$ and $\mathcal{A}^i_{\dddot{V}^{\mathcal{F}i}}$ remain fixed. And given that $\mathcal{A}^i$ is strictly positive by assumption, this also requires that T^i be chosen so that $\dddot{V}^{\mathcal{F}i} - V^i$ is unchanged which, given that $\dddot{V}^{\mathcal{F}i}$ is increasing in T^i, means that both $\dddot{V}^{\mathcal{F}i}$ and V^i must increase under this choice of T^i. But then $V^i_{T^i}$ must fall given the concavity of V^i, while $\dddot{V}^{\mathcal{F}i}_{T^i}$ must rise in light of the condition that $\dddot{V}^{\mathcal{F}i}$ is convex. And if $\dddot{V}^{\mathcal{F}i}$ is sufficiently convex, then the left-hand side of the top line of (45) must become negative under the hypothesized choice of T^i. But the right-hand side of the top line of (45) is λ, the Lagrange multiplier on the constraint in (37), which can never be negative. And so we may conclude that in response to the drop in τ from τ^N to τ^{PO}, the hypothesized increase in T^i would go too far to be consistent with optimal behavior on the part of the domestic government as characterized by the first-order conditions in (45). We have therefore established that a trade agreement that implements efficient politically optimal tariffs will increase the

[33] It is well known that under conditions of perfect competition the Johnson optimal tariff – which is the inverse of the foreign export supply elasticity – is weakly lower than the revenue maximizing tariff, implying that if the foreign export supply elasticity is low enough then removing this component from the tariff (as is implied in moving from the Nash to the politically optimal tariff) must reduce the tariff revenue collected by the government. We are also assuming that the share of tariff revenue that citizen i receives is exogenous to the setting of the tariff.

level of aggrievement in the domestic country (and hence, given our symmetry assumption, in each country) in this setting.

The intuition for this result is that, in the setting considered above, Nash international cost-shifting leads governments to be *overly* attentive to the aggrievement of their citizens in the Nash equilibrium, because part of the cost of their attentiveness is borne by citizens in the other country. When this international cost-shifting externality is addressed, as it will be by a fair trade agreement that implements politically optimal policies, the over-attentiveness of governments to the aggrievement of their citizens is eliminated, and the aggrieved citizens suffer higher (but now internationally efficient) levels of aggrievement. Hence, it is possible for the citizens of a country to feel that a trade agreement negotiated by their government has made the world more unfair for them even when their government is attentive to their fairness concerns.

We summarize with:

Proposition 6 *A trade agreement that is negotiated by governments that are attentive to the fairness concerns of their citizens may nevertheless heighten these concerns and lead to a greater level of aggrievement in each country than would prevail in the noncooperative Nash equilibrium. That is, relative to internationally efficient levels, Nash levels of aggrievement can either be too high or too low.*

We can also ask whether greater government attentiveness to fairness in terms of a higher γ and γ^* would lead a trade agreement between the governments to at least increase less the aggrievement of their citizens. Here it is easy to see that the answer is "Yes," for the simple reason that international cost-shifting in the Nash equilibrium is proportional to Nash trade volumes, and higher levels of γ and γ^* lead to higher Nash tariffs and lower Nash trade volumes, and hence smaller differences between Nash and politically optimal tariffs. In fact, if γ and γ^* are big enough, governments will choose tariffs in the Nash equilibrium that wipe out trade completely, and in that case these Nash choices will correspond to politically optimal tariffs and hence be internationally efficient.

6 Fairness for Others

Thus far, our analysis has focused on an individual's concerns regarding the fairness of his own position. However, Stantcheva (2023) has recently shown that when individuals are prompted to consider international trade, they express concerns not only about the distribution of benefits and costs in their own favor, but also a concern for others within the broader society. Gauthier (1986) and James (2014), as well as a significant body of philosophy literature on the subject, primarily address the fairness of trade in terms of the distribution of benefits across countries. The framework developed so far can be utilized to study these types of fairness concerns regarding the position of others, which may be motivated by altruism or a desire for distributive justice. We distinguish

between two possibilities. First, the "others" may be fellow citizens, potentially the poorest. Second, the "others" may be citizens of the partner country. In either case, in this section we now allow that citizens may have a taste for the utility of these "others."

A common feature in the literature on social preferences is that the material utility of others is incorporated directly into the material utility function of an individual i. Here, we can leverage the psychological dimension represented by the aggrievement function A^i defined in (9) by assuming that concern for others translates into vigilance regarding the fairness of their treatment. The perception of unfairness in the treatment of another individual elicits in individual i a psychological reaction analogous, in nature, to that elicited by the perception that he is being treated unfairly himself.

In this regard, our approach is close to that of Shaio (2009) and subsequent literature on social identification (e.g., Grossman and Helpman, 2021), which explicitly considers the psychological glow effect derived by an individual from the improvement of the position of other individuals in the social group to which he identifies.

Fairness towards fellow citizens We start with the case in which the "others" are i's fellow citizens. In order to focus on distributional concerns, we assume that citizen i has no concerns about procedural justice, only concerns about distributive justice. As we noted in section 3, this assumption implies that citizen i's fair utility for himself is a baseline fixed number. But since citizen i now may care not only about himself but also about others, we posit that his stringent claims are now represented by a *vector* of length N of fair utilities, with one level of fair utility for each domestic citizen (including himself), and with each of those levels being a fixed number. Let $\bar{\mathbf{V}}^{\mathcal{F}i}$ denote this vector, with its j^{th} element $V^{\mathcal{F}ij}$ denoting the utility that i deems fair for individual j, including himself ($j \in \{1, .., N\}$). Domestic citizen i's stringent claims are then given by

$$\bar{\mathbf{V}}^{\mathcal{F}i} \equiv \left\{V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}\right\}. \tag{46}$$

If citizen i has distributive concerns for citizen j, then the element $V^{\mathcal{F}ij}$ in the vector $\bar{\mathbf{V}}^{\mathcal{F}i}$ is strictly positive; if citizen i has no distributive concerns for citizen j, then the element $V^{\mathcal{F}ij}$ in the vector $\bar{\mathbf{V}}^{\mathcal{F}i}$ is equal to zero. The case where citizen i is self-centered is then a special case of (46) where $V^{\mathcal{F}ii} \geq 0$ and $V^{\mathcal{F}ij} \equiv 0$ for $i \neq j$.

Armed with citizen i's stringent claims function, we now proceed to formalize citizen i's level of aggrievement as before. If, in domestic citizen i's view, the material utility of any domestic citizen falls short of the level that citizen i deems to be fair for that citizen, then individual i is aggrieved. And, as in (10), aggrievement is commensurate with the difference, if positive, between fair utility and material utility. Thus, we extend (10) to include the N possible sources of aggrievement over distributive justice concerns, and define domestic citizen i's aggrievement function in the presence of fairness concerns toward

fellow citizens as follows:

$$A^i \equiv A^i(\max[0, V^{\mathcal{F}i1} - V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))], ... \\ ..., \max[0, V^{\mathcal{F}iN} - V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))]) \quad (47)$$

with A^i weakly increasing in each of its arguments.

The fair levels $V^{\mathcal{F}ij} \geq 0$ may capture different motivations behind i's social preference, such as those discussed in Section 2. For instance, $V^{\mathcal{F}i1} = ... = V^{\mathcal{F}iN} > 0$ indicates egalitarian preferences, while $V^{\mathcal{F}i1} > V^{\mathcal{F}i2} \geq 0$ might reflect i's idea that individual 1 deserves more than 2, based on principles of Contextual Justice or Desert.[34] The partial derivative of A^i with respect to its j^{th} argument reflects the intensity of i's aggrievement for the unfair situation of j, capturing how much i cares about the position of j. Thus, these partial derivatives parameterize feelings such as empathy or social identification. For instance, the partial derivative of A^i with respect to its j^{th} argument might be larger (or positive only) if j is a member i's in-group.

By (47) we can write

$$A^i \equiv \mathcal{A}^i(V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}, V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ..., V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))$$

with $\mathcal{A}^i$ weakly increasing in its first N arguments and weakly decreasing in its last N arguments. And building from (11), the total welfare function of domestic citizen i capturing his moral concerns for the utility of other fellow citizens is then given by:

$$W^i = V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)) \\ -\mathcal{A}^i(V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}, V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), .., V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)))$$

or

$$W^i \equiv \bar{W}^i(V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ... \\ ..., V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}) \quad (48)$$

with $\bar{W}^i$ weakly increasing it its first N arguments and weakly decreasing in its last N arguments. Similarly, building from (12), the total welfare function of a foreign citizen i capturing his moral concerns for the utility of his fellow citizens is given by:

$$W^{*i} \equiv \bar{W}^{*i}(V^{*1}(s^*, T^{*1}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ... \\ ..., V^{*N^*}(s^*, T^{*N^*}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), V^{\mathcal{F}*i1}, ..., V^{\mathcal{F}*iN}) \quad (49)$$

[34] The aggrievement function (47) does not account for inequity aversion or feelings of envy or spite. These sentiments imply that individual i might experience aggrievement even if $V^{ij\mathcal{F}} < V^j$ (say, because j is very rich). Despite their intriguing nature, we do not consider such sentiments here. In this respect, our approach aligns with that of Charness and Rabin (2002).

with $\bar{W}^{i*}$ weakly increasing it its first N^* arguments and weakly decreasing in its last N^* arguments.

Compared to the total welfare functions for domestic and foreign citizen i defined by (11) and (12), respectively, the material welfare levels entering as the first N arguments of (48) and the first N^* arguments of (49) have the same structure as the own-material welfare level that constitutes the first argument of (11) and (12). And since the stringent claims in (48) and (49) are constants, it is immediately apparent that domestic citizen i belongs to the set Ω_{Ia} defined in Section 5 while foreign citizen i belongs to the set Ω^*_{Ia}. According to Proposition 4, we may therefore state:

Proposition 7 *Concerns for distributive justice for one's fellow citizens do not by themselves change the purpose of a trade agreement, and hence these concerns by themselves do not have implications for the design of a fair trade agreement.*

Intuitively, the result reported in Proposition 7 reflects the fact that, while concerns for distributive justice for one's fellow citizens introduce new within-country externalities for each government to address, there is no new *international* externality created by these concerns, and hence no new problem for a trade agreement to solve.

Fairness towards foreign citizens We now consider the possibility that domestic citizen i is not only concerned about the distribution of income within his country, but also beyond the border. Now his stringent claim, denoted by $\hat{\mathbf{V}}^{\mathcal{F}i}$, is a vector of $N+N^*$ levels of fair utilities, which includes also the utilities deemed fair for foreign citizens,

$$\hat{\mathbf{V}}^{\mathcal{F}i} \equiv \left\{ V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}, V^{\mathcal{F}i1*}, ..., V^{\mathcal{F}iN^**} \right\}.$$

The aggrievement function, in addition to the N arguments capturing concern for fellow citizens (as in (47)), now also includes the N^* additional arguments $\max[0, V^{\mathcal{F}ij*} - V^{*j}(s^*, T^{*j}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*))]$ for $j = \{1, ..N^*\}$. These additional N^* arguments account for the aggrievement that would be triggered by the unfair treatment of foreign citizens, and they otherwise have the same interpretation as the first N arguments of the aggrievement function as discussed above.

Following the same steps as above, the total welfare of domestic citizen $i \in \{1, ..., N\}$ now becomes

$$\begin{aligned} W^i \equiv\ & \hat{W}^i(V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ..., V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \\ & V^{*1}(s^*, T^{*1}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ..., V^{*N^*}(s^*, T^{*j}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \\ & V^{\mathcal{F}i1}, ..., V^{\mathcal{F}iN}, V^{\mathcal{F}i1*}, ..., V^{\mathcal{F}iN^**}), \end{aligned} \tag{50}$$

with $\hat{W}^{*i}$ weakly increasing it its first $N+N^*$ arguments and weakly decreasing in its last $N+N^*$ arguments. And analogously, the total welfare of foreign citizen

$i \in \{1, .., N^*\}$) is given by

$$\begin{aligned} W^{*i} \equiv\ & \hat{W}^{*i}(V^{*1}(s^*, T^{*1}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), ..., V^{*N^*}(s^*, T^{*N^*}, p^*(\tau^*, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \\ & V^1(s, T^1, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), .., V^N(s, T^N, p(\tau, \tilde{p}^w), \tilde{p}^w(s, \tau, s^*, \tau^*)), \\ & V^{\mathcal{F}*i1*}, ..., V^{\mathcal{F}*iN^**}, V^{\mathcal{F}*i1}, ..., V^{\mathcal{F}*iN}). \end{aligned} \tag{51}$$

with $\hat{W}^{*i}$ weakly increasing it its first $N^* + N$ arguments and weakly decreasing in its last $N^* + N$ arguments.

Since aggrievement in one country now depends on the shortfall of material utility of citizens in the other country relative to the level that is deemed to be fair by citizens of the first country, and since the policies of the other country determine the material utility of each citizen in that country, the policies of the other country will now directly contribute to the level of aggrievement in the first country. Hence, the total welfare functions for domestic citizen i and foreign citizen i defined in (50) and (51) respectively now depend directly on the policy choices in the other country, similar to the case that arises for the respective sets Ω_{II} and Ω^*_{II} defined in Section 5, but now due to the dependence of aggrievement rather than stringent claims themselves on the policies of the other government.

Arguments analogous to those leading up to Proposition 4 and made for the respective sets Ω_{II} and Ω^*_{II} then allow us to conclude the following:

Proposition 8 *The purpose and design of a trade agreement is altered by fairness considerations when those considerations extend beyond the border to include distributive justice concerns for citizens of other countries. In this setting, deep-integration agreements over behind-the-border policies are required to reach the international efficiency frontier.*

In terms of our Figure 2 taxonomy, Proposition 8 implies that fairness considerations that extend beyond the border fall in the bottom left box of the figure. The novel international externality that drives the result reported in Proposition 8 also raises the possibility that a fair trade agreement could lead to lower market access and trade volumes, much as in the result reported in Proposition 2. This would happen if the negative terms-of-trade externality associated with a slight rise in a country's tariff beginning from Nash were outweighed by a positive international externality associated with the distributional consequences of the tariff increase for the tariff-increasing country.

An interesting special case of the result reported in Proposition 8 arises if it is accepted that cross-border distributive justice concerns typically extend in only one direction, from citizens of rich countries to the citizens of poor countries. When this is the case, the novel international externality created by cross-border distributive justice concerns extends in only one direction, from poor country policy choices to rich country citizens, and achieving the international efficiency frontier therefore only requires deep integration initiatives in one direction. That is, we may state the following:

Corollary *If rich country citizens have distributive justice concerns for the citizens of poor countries but poor country citizens do not have such concerns*

for the citizens of rich countries, then only the behind-the-border policies of poor countries need be the subject of international negotiation; shallow integration over rich-country policies is sufficient to achieve the international efficiency frontier in this setting.

Finally, notice that the need for deep integration in the presence of cross-border distributive justice concerns identified in Proposition 8 arises only because the citizens of one country care about the income distribution in another country *and* do not share the distributive goals of the government of the other country. If the distributive justice concerns were restricted to cross-country income differentials alone, as is often emphasized by the literature on fairness in trade (e.g., Gaulthier, 1986, and James, 2014) and as would be the case if domestic citizens sought a more equitable cross-country distribution of income but had trust in the foreign government to redistribute income within its country in a way that domestic citizens would approve of, then a shallow approach to integration continues to be a viable method for reaching the international efficiency frontier. This follows immediately, since the unilateral policy adjustments allowed under shallow integration can only enhance the ability of the policy-adjusting government to achieve its distributive goals – which citizens of the other country now share by assumption – while leaving the material welfare of each citizen of the other country unchanged.[35]

7 Conclusion

How do government concerns for fairness impact the purpose of a trade agreement? And how do those concerns impact the implied design features of a trade agreement that can serve this purpose? Taking a novel bottom-up approach where government concerns for fairness derive from the moral sentiments of their constituents, we have shown how formal answers to these questions can be provided, and we have shown that the answers hinge on *how* a country's citizens evaluate whether they have been treated fairly by trade. Accordingly, our findings point to a detailed understanding of real-world perceptions of trade-related fairness concerns as the key input into the appropriate design of fair trade agreements. And our findings suggest that, as currently designed, the GATT/WTO is well-equipped to allow its member governments to address many, though not all, of the possible trade-related fairness concerns of their citizens.

Relative to a top-down approach to concerns for fairness, our bottom-up approach also reveals an important broader insight. Under a top-down view of

[35] The viability of a shallow agreement in these circumstances does hinge on the assumption that countries' concern for others comes from a position of altruism. If instead countries were inequity averse or motivated by spite or envy (see also note 34), then if allowed to make unilateral behind-the-border policy choices subject to a market-access preservation rule of the kind considered in section 4, at least one government would act in a manner that is not in the interest of the other country's citizens, thereby exacerbating the novel international externality problem and undermining the ability of shallow integration to reach the international efficiency frontier.

fairness, countries must first agree on a moral standard by which to judge fairness before they could incorporate fairness concerns into their trade agreements. But our bottom-up approach reveals that, even with differences of opinion within and across countries over what is the "right" moral standard by which fairness should be judged, countries can still benefit from cooperating over fairness concerns if they negotiate to internalize the international externalities created by the moral sentiments of their citizens.

Throughout the paper we have focused on the possibility that fairness concerns might change the purpose of a trade agreement. And where we have found that the purpose would change, we have emphasized the possibility that the utility of key design features of the GATT/WTO, such as its shallow approach to integration and its emphasis on tariff bindings as the central negotiated legal commitment, might then be undermined. What we have not considered above is the question whether the design of the GATT/WTO could itself also be said to conform with top-down views of what a fair international trade institution should deliver. We close by briefly considering some of these views and suggesting one possible answer to this question. Our purpose is both to illustrate the kinds of thorny normative questions that our bottom-up analysis has been able to sidestep, and to point to particular design features of the GATT/WTO that, from the perspective of the top-down literature on fairness in trade, may be worthy of further study.

As we noted in section 2, much of the literature on fairness in trade emphasizes the division of the gains from trade across countries. In contrast to our approach, this literature typically abstracts from the individual's point of view, striving instead for a definition of fairness that can be universally applied to all countries and individuals. A prominent paper in this literature is James (2014), who describes the subject of fairness in trade in these general terms:

> The basic subject of fairness in trade is an *international social practice of market reliance*, that is to say, a social practice in which countries mutually rely on common markets (in goods, services, or capital) for the sake of augmenting their national incomes – what Adam Smith famously called the "wealth of nations." ... The international practice of market reliance can be organized in different ways, with varying consequences for the national incomes of different countries and for the socio-economic prospects of their respective social classes. ... The collective choice of organization, through negotiated agreements or trend-setting unilateral action, is therefore subject to basic moral demands of fairness, beyond mere considerations of national interest, efficiency, or overall welfare.
>
> Chief among them are requirements of *structural equity*, which concern how the trade practice distributes the benefits and burdens it creates, among different countries, and among their respective classes, according to principles that no one can reasonably complain of. To say that the international market reliance practice has an equitable structure, or is structurally equitable, is to say that it

> distributes the benefits and burdens it creates according to a pattern that is reasonably acceptable to every country and class affected. (pp 178-179, emphasis in the original, footnotes omitted)

James advocates a notion of structural equity that can be boiled down to three principles. Two of these principles are concerned with structural equity within each country. The third relates to structural equity between countries in a trade agreement, and it is this third principle that James takes as the main focus of structural equity in the context of trade agreements.[36] This principle is stated as follows:

> *International Relative Gains*: national income gains due specifically to international trade are to be distributed equally, unless greater gains flow (e.g. via special trade privileges) to poor countries. (James, 2014, p 181, footnotes omitted)

As Haubermann (2017) observes, in taking this stand James (2014) is rejecting the earlier reasoning of Gauthier's (1986) "contribution-based proposal of gain distribution":

> According to Gauthier, each participating country would try to claim as much of the gains of international trade as possible, and the distribution would eventually be settled by rationally self-interested bargaining. This way of distribution seems morally inadmissible, since no country has a morally relevant interest in the whole of the gains, but each participating country's interest in greater rather than lesser shares at most amounts to a presumptive claim to equal gains. (Haubermann, 2017, p 7)

While Gauthier describes a "power-based" view of trade bargaining, James seems to advocate a "rules-based" concept of fairness in trade, arguing that the requirement of structural equity then demands that the gains from trade should be distributed equally across all countries. Of course, whether this is a principle of fairness in trade "that no one can reasonably complain of" is in the eye of the beholder. A different rules-based concept of fairness in trade is put forward by Risse and Wollner (2019), who equate trade injustice with the concept of exploitation, and define exploitation as unfairness through power:

[36] As James (2014) notes:

> A crucial feature of the account is that it is *international* in a qualified yet particularly strong sense: it provides no scope for comparing levels of gain for any two individuals of the world. The first principle does consider harm to individuals (or members of social classes), whether or not they live within the trading system. But as far as the *gains* of the system are concerned, comparisons between individuals are only allowed (under the second principle) *within* a given trading society. Assuming no one is harmed, the distribution of gains across societies is evaluated (under the third principle) at the level of whole countries. This is despite the fact that we can easily imagine a further, specifically "global" or "cosmopolitan" fairness principle that directly limits the relative gains of any two individuals of the world. (p 181, emphasis included in the original)

> Exploitation, on our general ecumenical account, is unfairness through power. The version that applies to trade characterizes exploitation as power-induced failure of reciprocity. Importantly, non-individual actors like groups or institutions may exploit or be exploited. (Risse and Wollner, 2019, p 14)

As we have throughout the paper, we avoid taking a stand here on which concept of fairness in trade best embodies a principle "that no one can reasonably complain of." Instead, here we simply offer an additional concept for consideration, one that combines elements of Gauthier (1986), James (2014) and Risse and Wollner (2019) and which offers a particular interpretation of the concept put forward by Risse and Wollner, and one that, as we note below, GATT/WTO rules appear to be reasonably well-equipped to deliver. The principle we put forward is this: All countries should agree to adopt the policies that they would adopt if they ignored their power to influence the terms of trade with their policy choices; that is, a fair trade agreement should implement the political optimum.

As Bagwell and Staiger (1999, 2001, 2002) have shown in the absence of fairness considerations, if all countries were to agree to their politically optimal policies, they would achieve a point on the international efficiency frontier. And importantly for our discussion here, it can be argued that this particular point has a claim to fairness, in that it is a point on the international efficiency frontier that is defined without regard to the power possessed by individual countries to threaten adverse terms-of-trade consequences for their trading partners in order to achieve for themselves more favorable bargaining outcomes on the frontier. For example, if all governments sought to maximize their national income with their tariff choices, then the political optimum would correspond to reciprocal free trade, corresponding to a point on the efficiency frontier for that case which would not reflect the ability of countries whose tariffs have greater impacts on the terms of trade to move the outcome of the tariff bargain in a direction that was more favorable to them. In this sense, one could say that movements to points away from the political optimum represent what Risse and Wollner (2019) call exploitation, or unfairness through power, where power here has a specific meaning: the ability to use (or to threaten to use) one's policies to adversely impact the terms of trade of one's trading partners.

Finally, we note that the political optimum is a point on the international efficiency frontier that GATT/WTO rules seem particularly well-suited to deliver. In particular, as discussed in Bagwell and Staiger (2016), under the GATT/WTO pillars of reciprocity and nondiscrimination as embodied in the most-favored-nation (MFN) principle, the political optimum may be viewed as a possible focal outcome for the rounds of multilateral tariff negotiations that have delivered the core accomplishments of GATT. Of course, the efficiency properties of the political optimum were derived by Bagwell and Staiger in a world without fairness considerations, and to the extent that fairness considerations introduce new international externalities as we have characterized those possibilities above, the efficiency of the political optimum will be disrupted and

the attractiveness of our proposed principle of fairness in trade will be diminished. But as our results above suggest, there are many circumstances where concerns for fairness will not change the purpose of a trade agreement and the political optimum remains efficient. To the extent that those circumstances describe important features of the real world, our proposed principle of fairness in trade seems attractive. In this light, the particular design features of the GATT/WTO that have been shown to guide outcomes toward the political optimum may, from the perspective of the top-down literature on fairness in trade, be worthy of further study.

8 References

Abbott, Kenneth W. 1996. "Defensive Unfairness: The Normative Structure of Section 301," Chapter 9 in Jagdish Bhagwati and Robert E. Hudec (eds) *Fair Trade and Harmonization*, Volume 2. MIT Press, Cambridge MA.

Arrow, Kenneth J. 1973. "Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice," *Journal of Philosophy* 70(9): 245-63.

Bagwell, Kyle and Robert W. Staiger. 1999. "An Economic Theory of GATT," *American Economic Review* 89(1): 215-248.

Bagwell, Kyle and Robert W. Staiger. 2001. "Domestic Policies, National Sovereignty, and International Economic Institutions," *Quarterly Journal of Economics* 116: 519-562.

Bagwell, Kyle and Robert W. Staiger. 2002. *The Economics of the World Trading System*, Cambridge: The MIT Press.

Bagwell, Kyle and Robert W. Staiger. 2012. "The economics of trade agreements in the linear Cournot delocation model," *Journal of International Economics* 88: 32-46.

Bagwell, Kyle and Robert W. Staiger. 2016. "The Design of Trade Agreements," in Kyle Bagwell and Robert W. Staiger (eds) *The Handbook of Commercial Policy*, vol 1A, December.

Bhagwati, Jagdish. 1996. "The Demands to Reduce Domestic Diversity among Trading Nations," Chapter 1 in Jagdish Bhagwati and Robert E. Hudec (eds) *Fair Trade and Harmonization*, Volume 1. MIT Press, Cambridge MA.

Bhagwati, Jagdish and Robert E. Hudec (eds). 1996. *Fair Trade and Harmonization*, Volumes 1 & 2. MIT Press, Cambridge MA.

Binmore, Ken. 1994. *Game Theory and the Social Contract, Vol 2: Just Playing*. Cambridge, MA: MIT Press.

Brown, Andrew G. and Robert M. Stern. 2007. "Concepts of Fairness in the Global Trading System," *Pacific Economic Review* 12(3): 293-318.

Buchanan, James M. 1986. *Liberty, Market and State: Political Economy in the 1980s*. NY: NYU. Press, Columbia U. Press.

Cass, Ronald A. and Richard D. Boltuck. 1996. "Antidumping and Countervailing-Duty Law: The Mirage of Equitable International Competition," Chapter 8 in Jagdish Bhagwati and Robert E. Hudec (eds) *Fair Trade and Harmonization*, Volume 2. MIT Press, Cambridge MA.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3): 817–869.

Davidson, Carl, Steve Matusz, and Doug Nelson. 2006. "Fairness and the Political Economy of Trade," *The World Economy* 29(8): 989–1004.

Di Tella, Raphael and Dani Rodrik. 2020. "Labour Market Shocks and the Demand for Trade Protection: Evidence from Online Surveys," *The Economic Journal* 130(628): 1008–1030

Fehr, Ernst, Oliver Hart and Christian Zehnder. 2011. "Contracts as Reference Points – Experimental Evidence," *American Economic Review* 101(2): 493-525.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economic*s 114(3): 817-868.

Gauthier, David. 1986. *Morals by Agreement*, Oxford: Oxford University Press.

Grossman, Gene M. and Elhanan Helpman. 2021. "Identity Politics and Trade," *Review of Economic Studies*. 88: 1101–1126.

Guriev, Sergei and Elias Papaioannou. 2022. "The Political Economy of Populism," *Journal of Economic Literature* 60(3): 753–832.

Haidt, Jonathan. 2012. *The Righteous Mind,* Penguin Books.

Harsanyi, John C. 1975. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review* 69: 594-606.

Hart, Oliver, and John Moore. 2008. "Contracts as Reference Points," *Quarterly Journal of Economics*, 123(1): 1–48.

Haubermann, Johann Jakob. 2017. "Fairness in international trade policy: equality and differential treatment in theory and practice," Free University of Berlin Working Paper, October.

Hudec, Robert E. 1996. "Introduction to the Legal Studies," in Jagdish Bhagwati and Robert E. Hudec (eds) *Fair Trade and Harmonization*, Volume 2. MIT Press, Cambridge MA.

Inglehart, Ronald F. and Pippa Norris. 2016. "Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash," Faculty Research Working Paper Series RWP16-026, August: Harvard Kennedy School.

James, Aaron. 2005. "Distributive Justice without Sovereign Rule: The Case of Trade," *Social Theory and Practice* 31(4): 533-559.

James, Aaron. 2014. "A Theory of Fairness in Trade," Moral Philosophy and Politics, 1(2): 177–200.

Kahneman, Daniel, Jack L. Knetsch and Richard Thaler. 1986. "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* 76: 728-41.

Kapstein, Ethan B. 2008. "Fairness Considerations in World Politics: Lessons from International Trade Negotiations," *Political Science Quarterly* 123(2): 229-245.

Konow James. 2003 "Which Is the Fairest One of All? A Positive Analysis of Justice Theories," *Journal of Economic Literature* 41(4):1188-1239.

Kurjanska, Malgorzata and Mathias Risse. 2008. "Fairness in trade II: export subsidies and the Fair Trade movement," *Politics, Philosophy & Economics* 7(1): 29-56.

Kuziemko, Ilyana, Nicolas Longuet-Marx and Suresh Naidu. 2023. "Compensating the losers? Economic Policy and Partisan Realignment in the US," NBER Working Paper No 31794, October.

Meyer, Timothy. Forthcoming. "Consumption Governance: The role of production and consumption in international economic law," *Brigham Young University Law Review*.

Miller, David. 1992. "Distributive Justice: What People Think," *Ethics* 102(3): 555–93.

Miller, David. 2017. "Fair Trade: What Does It Mean and Why Does It Matter?," *Journal of Moral Philosophy* 14: 249-269.

Narlikar, Amrita. 2006. "Fairness in International Trade Negotiations: Developing Countries in the GATT and WTO," *The World Economy*: 1005-1029.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. NY: Basic Books.

Passarelli, Francesco and Guido Tabellini. 2017. "Emotions and Political Unrest," *Journal of Political Economy* 125(3): 903-946.

Pauwelyn, Joost and Charlotte Sieger-Gasser. 2024."Addressing Negative Effects of Trade Liberalization: Unilateral and Mutually Agreed Flanking Policies," mimeo, Geneva Graduate Institute.

Porter, Eduardo. 2024. "Despite Biden's efforts, working-class voters still don't trust Democrats," Opinion Article in *The Washington Post*, March 11.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Belknap Press of Harvard U. Press.

Risse, Mathias. 2007. "Fairness in trade I: obligations from trading and the Pauper-Labor Argument," *Politics, Philosophy & Economics* 6(3): 355-377.

Risse, Mathias and Gabriel Wollner. 2019. *On Trade Justice: A Philosophical Plea for a New Global Deal*, Oxford University Press: United Kingdom.

Rodrik, Dani. 2021. "Why Does Globalization Fuel Populism? Economics, Culture, and the Rise of Right-Wing Populism," *Annual Review of Economics* 13: 133-170.

Roth, Alvin E. 2007. "Repugnance as a Constraint on Markets". *Journal of Economic Perspectives* 21(3): 37–58.

Shaffer, Gregory. 2024. "Package Treaties: Addressing the Negative Effects of Trade," mimeo, Georgetown University Law Center.

Shayo, Moses. 2009. "A Model of Social Identity with an Application to Political Economy: Nation, Class, and Redistribution," *American Political Science Review* 103: 147–174.

Staiger, Robert W. 2022. *A World Trading System for the Twenty-First Century*, MIT Press, Cambridge MA.

Staiger, Robert W. and Alan O. Sykes. 2011. "International Trade, National Treatment, and Domestic Regulation," *Journal of Legal Studies* 40(1): 149-203.

Staiger, Robert W. and Alan O. Sykes. 2021. "The Economic Structure of Trade-in-Services Agreements," *Journal of Political Economy* 129(4): 1287-1317.

Stantcheva, Stephanie. 2023. "Understanding of Trade," mimeo, Harvard University, August 31.

Winkelmann, Liliana and Rainer Winkelmann. 1998. "Why are the Unemployed So Unhappy? Evidence from Panel Data," *Economica* 65(257): 1–15.