Domestic Policies in Self-Enforcing Trade Agreements*

Philip Sauré†
Swiss National Bank
May 3, 2012

Abstract

If all cross-country externalities travel through the terms-of-trade, efficient trade agreements target the terms-of-trade but ignore domestic policies. This argument has been advanced by prominent studies on trade agreements. The present paper shows that its logic fails under dynamic production possibilities – for example, if factor accumulation takes time. In these cases, past policies shape current production possibilities and thus affect defection temptations. Therefore, self-enforcing trade agreements that leave the choice of domestic policies to individual countries risk that countries abandon the zone of voluntarily cooperation while optimizing their policies. Avoiding this from happening inflicts inefficiencies on the world economy that are absent in trade agreements that target policies directly. The losses are strictly positive except for knife-edge cases, which existing studies have focussed on.

*The views expressed in this paper are the author’s views and do not necessarily represent those of the Swiss National Bank
†Email: philip.sauere@snb.ch.
1 Introduction

If all cross-border distortions and inefficiencies travel through international prices, trade agreements need only address terms-of-trade. Based on this logic, the literature on trade agreements has advised trade negotiators to ignore domestic policies and simply negotiate "market access," i.e., the terms-of-trade. In short, efficient trade agreements are terms-of-trade agreements only.

The present paper shows that this logic fails under two natural conditions: if economic conditions are linked intertemporally and trade agreements are constrained by the requirement of self-enforceability. Under these conditions, a country’s policy mix can affect its defection temptation while leaving the terms-of-trade unchanged. Therefore, a country that optimizes its policy mix within the confines of a terms-of-trade agreement may abandon the zone in which it voluntarily complies to the trade agreement and consequently defect.

Within a general framework, I analyze conditions under which terms-of-trade agreements are efficient and show that these conditions are met only in knife-edge cases, which previous studies have focussed on. Generally, as long as self-enforcement constraints are binding, terms-of-trade agreements are inefficient, inflicting losses on the global economy that are absent in trade agreements directly targeting tariffs and domestic policies.

The present paper’s insight relies on one crucial assumption, which existing work on trade agreements has implicitly ruled out: economic conditions are intertemporally linked such that past policies affect present economic conditions. For example, an economy’s past investment subsidy affects its present capital stock. This almost pathetically basic assumption has been neglected by previous studies but suffices to disprove their findings on the efficiency of terms-of-trade agreements. As the past policy mix influences present production possibilities, it affects defection temptation and must therefore be targeted in optimal self-enforcing trade agreements.

The natural assumption of intertemporal linkages challenges the conventional wisdom concerning trade agreements. The exact logic of my critique relies on the interaction between intertemporal economic linkages and the temptation to defect on trade agreements. While the former scarcely requires any motivation, the latter needs a word of explanation.

In line with existing work, I adopt the view that, in the absence of a central author-
ity to enforce international contracts, voluntary enforcement is a central prerequisite of international cooperation.¹ Fundamental to the concept of self-enforcing trade agreements is the optimal tariff argument, according to which large countries are tempted to exert their market power on international goods markets by unilaterally charging tariffs, thereby distorting the terms-of-trade in their favor. Thus, starting from zero trade barriers, each country has an incentive to unilaterally raise tariffs. If all countries do so, however, the resulting equilibrium is characterized by high tariffs and substantial efficiency losses for the world economy. This situation can be referred to as a trade war.

Because countries trade repeatedly, the inefficiencies of a trade war can be alleviated by trade agreements in the spirit of repeated games. In such trade agreements, the temptation to distort terms-of-trade remains because a country’s unilateral defection would yield benefits for a transition period in which trade partners cannot react. After such a transitory period of defection, however, a trade war erupts, the efficiency losses of which constitute the costs of defection. Self-enforcing trade agreements ensure that, for each country, the transitory gains from defection fall short of future losses from a trade war. While reducing distortions, the defection temptation of countries must be checked by self-enforcement constraints.

The present paper argues that past policies influence present economic conditions and that, therefore, the policy mix of a trade agreement becomes relevant for defection values and thus for self-enforcement constraints. Hence, efficient trade agreements directly target a country’s exact policy mix. Implicit in this last logical step is the view that trade agreements must explicitly state the exact limitations of a country’s policies. One may, of course, object that countries voluntarily refrain from setting policies that are not self-enforcing. In fact, any move out of the zone of self-enforcing policy mixes would lead to a breakdown of the agreement because trade partners anticipate imminent defection. Therefore, the policy mix is tacitly complied with and again, targeting the terms-of-trade in the trade agreement should be enough to obtain efficiency. Indeed, in the tradition of tacit collusion, no policies must be formally fixed by a trade agreement if all players know each other’s objectives, defection temptations and bargaining power. Thus, the logic of this objection applies not only to the exact combination of policies but to the entire agreement,

¹This interpretation of trade agreements goes back to Yarbrough and Yarbrough (1986) and Dixit (1987).
including the terms-of-trade. Cooperation should be possible without any formal contracts to determine either policies or the terms-of-trade. However, if anything must be laid down in the legal text of a trade agreement, then the exact mix of domestic policies is just as relevant as the terms-of-trade.

The modeling framework utilized in this paper is very general, incorporating those in Ederington (2001) and Bagwell and Staiger (2001) as special cases. Despite this general framework, I nevertheless follow the standards of the literature by assuming that all international externalities travel through the terms-of-trade. This assumption restricts my analysis to a setup within which previous work has shown that negotiating market access removes all cross-border inefficiencies.

Since the work of Yarbrough and Yarbrough (1986) and Dixit (1987), trade theory has generally understood trade agreements as a set of rules that encourage trade integration but, in the absence of a supra-national executive authority, must be self-enforcing. An important body of literature has built on this concept. Some studies analyze the effects of adjustment costs, which generally make the consequences of a defection on trade agreements harsher and defection less attractive. Adjustment costs can generate endogenous gradualism in trade liberalization when output or production possibilities change gradually (Staiger (1995), Devereux (1997) and Furusawa and Lai (1999)). In addition, differences in country size can induce dynamic tariff reductions in an otherwise static economic environment (Park (2000) and Bond and Park (2002)). Furthermore, adjustment costs induce endogenous shifts in bargaining positions, thus generating a version of a hold-up problem (see McLaren (1997)).

Another important body of literature analyzes how to address multiple policy instruments and trade agreements. The leitmotif of this literature is that national governments should be free to choose their preferred mix of domestic policies. Ederington (2001) finds that within self-enforcing trade agreements "the most efficient means of countering the incentive to deviate from the agreement is to relax only trade policy." In the same spirit, Bagwell and Staiger (2001) state that "the inefficiency associated with unilateral policy choices reflects a problem with the market access, not with the policy mix," while Bagwell and Staiger (2006) argue that the direct restrictions of production subsidies within the "WTO subsidy rules interfere with the ability of governments to structure their tariff
negotiations so as to achieve efficient policy combinations." The present paper shows that such assertions concerning the efficiency of trade agreements apply under a set of specific assumptions, which almost surely do not hold.

The remainder of the paper is organized as follows. Section 2 gives a graphical representation of the conditions under which terms-of-trade agreements are efficient. Section 3 formalizes the diagrammatic analysis and shows that the identified conditions are almost surely violated. Section 4 concludes.

2 A Graphical Representation of the Argument

This section illustrates under which conditions terms-of-trade agreements – i.e. trade agreements that target only the terms-of-trade – are efficient.

I consider a world economy consisting of two countries. Both countries may set import tariffs $T$ and a generic policy $P$. The analysis will be limited to trade agreements that implement time-invariant cooperation policies. I take the perspective of the one country, called Home. The self-enforcement constraint of Homes is assumed to be binding. For the sake of simplicity, I neglect foreign defection temptation; i.e., I assume that the self-enforcement constraint for the foreign country is not binding. This assumption only holds if the foreign country discounts future benefits at very low rates.

Trade Agreements Targeting Policies – The Traditional View. Consider first a trade agreement addressing the policies $T$ and $P$. Figure 1 represents the starting point of a generic optimization problem of the trade agreement. The $(T, P)$ space comprises all possible combinations of Home’s policy instruments.

Social welfare can be thought of as a weighted sum of both countries’ welfare. Conditional on the successful implementation of a trade agreement involving Home’s respective policies, the level of social welfare is maximal at $(T^o, P^o)$ (the index $o$ stands for the unconstrained social optimum). The strategy that maximizes Home’s individual welfare is $(T^i, P^i)$ (the index $i$ stands for the unconstrained individual optimum). Both social and individual welfare are conditional on the successful implementation of the respective trade agreement. The social optimum $(T^o, P^o)$ and the individual optimum $(T^i, P^i)$ do
Figure 1: The \((T, P)\) space represents all possible combinations of Home’s policies. \((T^o, P^o)\) represents the social optimum and \((T^i, P^i)\) represents Home’s individual optimum. The dotted lines represent schedules of constant world prices; the black arrows indicate the direction of increasing individual welfare. Because foreign welfare is constant at constant world prices, Home’s and social welfare increase and decrease in parallel along the dotted lines.

not coincide, reflecting international conflicts of interest based on the usual terms-of-trade externalities.\(^2\)

Consider now, for any given world price \(\pi\), the combinations \((T, P)\) that leave \(\pi\) unchanged. These combinations are represented by the thin dotted lines. How does Home’s individual welfare change along the dotted lines? Figure 2 illustrates the answer to this question. On the dotted line cutting through \((T^i, P^i)\), individual welfare must increase with proximity to \((T^i, P^i)\) because individual welfare is maximal at \((T^i, P^i)\).\(^3\) The direction of increasing individual welfare is indicated by the small black arrows.

Clearly, on each dotted line, there is a maximum of Home’s welfare, which is indicated by the respective small circle. Connecting all the circles creates a schedule represented by the dashed line.

\(^2\)See the discussion of the optimal tariff argument above.
\(^3\)This statement holds at least locally.
Figure 2: The dashed line represents the schedule of maximal individual welfare for given world prices (along the dotted lines). As foreign welfare is constant for constant world prices, social welfare is maximal whenever Home’s welfare is. Therefore, \((T^o, P^o)\) lies on the dotted line.

Because all cross-border inefficiencies travel through the terms-of-trade and terms-of-trade are constant along the dotted lines, changes of Home policies along the dotted lines do not change the welfare of the foreign country. Therefore, as social welfare consists of Home’s and foreign welfare, movements along the dotted lines increase and decrease Home’s and social welfare in parallel. This observation implies that on each dotted line social welfare is maximal whenever Home’s welfare is maximal. Thus, the social optimum \((T^o, P^o)\) lies on the dashed line, i.e., the schedule of Home’s individual maxima.\(^4\)

Figure 3 adds to the picture the social and individual indifference curves in the neighborhood of the dashed line. Home’s welfare, represented by the grey solid lines, increases in the direction of \((T^i, P^i)\). Social welfare, represented by the black solid lines, increases towards \((T^o, P^o)\).

\(^4\)The observation also implies that \((T^o, P^o)\) and \((T^i, P^i)\) cannot lie on the same iso-price schedule.
Figure 3: Solid black lines indicate social indifference curves; solid grey lines indicate Home’s individual indifference curves. Both are tangent to the iso-price schedule along the dashed line connecting \((T^o, P^o)\) and \((T^i, P^i)\).

Within the diagrammatic framework laid down in Figure 3, I now turn to the graphical analysis of trade agreements. An unconstrained trade agreement implements the social optimum \((T^o, P^o)\). However, if Home’s self-enforcement constraint is binding this outcome is impossible. In this case, the optimal trade constrained agreement should, intuitively, implement policies that lie somewhere on the dashed line. Otherwise, a change to an alternative policy combination could improve social welfare, Homes’ welfare, or both. The exact location depends on the tightness of the constraints.

Figure 4 incorporates Home’s self-enforcement constraint. The domestic self-enforcement constraint defines a subset of the \((T, P)\)-plane, defined as the set of all self-enforceable combinations of policies \(T\) and \(P\). This subset is represented by the grey area. Reflecting the assumption that the social optimum is not enforceable, \((T^o, P^o)\) lies outside the grey area. The uncooperative domestic policies \((T^i, P^i)\) are trivially sustainable; hence, \((T^i, P^i)\) lies
Figure 4: The grey area represents the combinations of Home’s self-enforcing policies. Here, the border of this area coincides with one of Home’s indifference curves. The constrained optimal trade agreement implements the domestic policies \((T^c, P^c)\). A trade agreement targeting \(\pi^c\) only makes Home choose \((T^c, P^c)\) and is self-enforcing.

inside the grey area. In Figure 4 the borders of the grey area, i.e., the schedule at which the self-enforcement constraint is marginally binding, coincides with Home’s indifference curve. Again, this is not a coincidence but a reflection of standard modeling assumptions, which will be discussed in detail in Section 3. Taking this feature as given, the constrained optimal trade agreement implements Home’s policies \((T^c, P^c)\). It thereby maximizes world welfare within the grey area, indicating that the agreement is self-enforceable.

**Terms-of-Trade Agreements – The Traditional View.** With the help of Figure 4, one can also analyze terms-of-trade agreements, i.e., the class of trade agreements that do not directly target the policies \(T\) and \(P\) but rather the world price \(\pi\). In particular, let \(\pi^c\) be the world price corresponding to Home’s policies \((T^c, P^c)\). Consider the trade agreement

\(^5\)Again, I neglect foreign incentives to defect and thus the possibility that such a strategy would trigger a trade war.
Figure 5: The grey area represents the mix of Home’s policies that is self-enforcing. The border of this area does not coincide with the domestic indifference curves. The constrained optimal trade agreement implements the domestic policies \((T^c, P^c)\) with the corresponding world price \(\pi^c\). Instead, a trade agreement targeting \(\pi^c\) will make Home choose the policies \((T^d, P^d)\) are not self-enforcing.

that fixes the world price at \(\pi^c\) but allows Home free choice of its policy mix \((T, P)\), only requiring that the resulting world price be \(\pi^c\). In other words, Home’s policies are bound to the dotted line crossing \((T^c, P^c)\). If Home optimizes under this constraint and complies with the trade agreement, it will choose \((T^c, P^c)\) as its individually optimal policy mix. By construction, \((T^c, P^c)\) is also socially optimal and self-enforcing.

**Trade Agreements Targeting Policies – The General View.** From a naive graphical perspective, the condition that the borders of Home’s self-enforcement constraint coincide (locally) with its indifference curve seems very restrictive. To consider a more general case, Figure 5 illustrates a situation in which the schedule at which the self-enforcement constraint is binding does not coincide with Home’s indifference curve. In this case, the
Figure 6: Home’s individual welfare along the iso-price schedule \( \pi^c \). The dotted line represents welfare under cooperation, and the thick dashed line represents welfare under defection on the trade agreement \( (T(P), P) \) given that economic conditions have reached their steady-state \( Z_\infty \). The thin dotted line represents Home’s welfare if economic conditions have been established at \( Z_\infty(T^c, P^c) \) and Home changes to the policies \( (T(P), P) \) and then defects. Overall, policies are self-enforcing if \( P \) lies to the left of \( P^c \).

But why does Home not simply switch to the policies \( (T^d, P^d) \), increase its welfare and defect? The answer to this question is related to the assumptions the present paper adds to the standard framework: economic conditions are intertemporally linked and change sluggishly. In particular, under an established trade agreement with policies \( (T, P) \), economic conditions converge to their steady state, which I indicate by \( Z_\infty \). In general, these economic conditions will depend on Home’s precise policy mix, so that one can write \( Z_\infty(T, P) \).

By changing the economic conditions, a change in \( Z_\infty \) impacts the value of defection. As will become clear in the next section, the value of defection will generally depend on the policies \( (T, P) \) through \( Z_\infty \) but not on the policies directly. Such a case is illustrated in Figure 6.
Figure 6 illustrates the values of cooperation and defection along constant terms-of-trade $\pi^c$, i.e., the dotted line through $(T^c, P^c)$ and $(T^d, P^d)$ in Figure 5. Let $T(P)$ be the function that describes this schedule. The horizontal axis in Figure 6 represents Home’s policy $\pi$. On the vertical axis is Home’s welfare $V$, as a function of the steady-state $Z_\infty(T(P), P)$. The thick bold line and the dashed line are the values of cooperation and defection, respectively, given the steady-state value of economic conditions, $Z_\infty(T(P), P)$.

Conditional on cooperation, Home’s welfare is maximized at $P^d$ (compare 5). This policy is not self-enforcing, however, because the value of defection falls short of the value of cooperation only to the left of $P^c$.

What will happen, if Home defects on the trade agreement $(T^c, P^c)$? Because the economic conditions $Z$ do not change instantaneously, defection on the trade agreement $(T^c, P^c)$ does not render the high welfare as a defection on a trade agreement $(T^d, P^d)$. Instead, the corresponding welfare remains at the lower value $V[Z_\infty(T^c, P^c)]$, which is illustrated by the thin dashed line in Figure 6. Therefore, Home cannot not increase its individual welfare by switching to the policies $(T^d, P^d)$ and subsequently defect. The policies $(T^c, P^c)$ are self-enforcing.

**Terms-of-Trade Agreements – The General View.** Finally, consider again the type of trade agreement that does not target the policies $T$ and $P$ but rather the world price $\pi$. Once again, let $\pi^c$ be the world price corresponding to Home’s policies $(T^c, P^c)$ and consider a trade agreement that fixes the world price $\pi^c$, otherwise leaving Home free to choose its policy mix $(T, P)$. Under this trade agreement, Home’s policies can move along the dotted line to the individually optimal policy mix, $(T^d, P^d)$, where Home defects on the trade agreement. In sum, the optimal trade agreement is not self-enforceable.

Of course, one may object that Home will not choose to play $(T^d, P^d)$, knowing that such strategies are not sustainable and will therefore lead to a breakdown of the agreement before implementation. For this reasons, targeting the terms-of-trade $\pi$ in the trade agreement should be sufficient to obtain efficiency. In principle, this objection is correct. Indeed,
Figure 7: The grey area represents the mix of Home’s self-enforcing policies. The border of this area does not coincide with the domestic indifference curves. A trade agreement targeting only the terms of trade cannot implement the domestic policies \((T^c, P^c)\). The globally optimal terms-of-trade agreement fixes the world price \(\pi^{cc}\) instead, which results in the equilibrium policies \((T^{cc}, P^{cc})\). There are, however, global efficiency losses relative to the trade agreement implementing the policies \((T^c, P^c)\).

very much in the tradition of tacit collusion, nothing needs to be formally fixed by a trade agreement if all players know each other’s objectives, defection temptations and bargaining power. However, the logic of this objection applies not only to the exact combination of policies \((T, P)\) but also to the terms-of-trade \(\pi\). If the constrained social optimum can be attained tacitly, i.e., without any formal contract determining \(T^c, P^c\) or \(\pi^c\). However, if anything must be laid down in the legal text of a trade agreement, then the precise mix of domestic policies is just as relevant as the terms-of-trade.

Finally, one may wonder which terms-of-trade agreement is sustainable if that which implements \(\pi^c\) is not. The answer to this question is illustrated in Figure 7. The self-enforcing terms-of-trade agreement must target a world price \(\pi\) that guarantees that Home’s movement along the iso-price line to its individual optimum does not bring it out of the
grey area where its self-enforcement constraint is satisfied. Clearly, among those terms-of-trade agreements satisfying this requirement, that which targets the world price $\pi^{cc}$ maximizes social welfare. Home’s resulting effective policies are $(T^{cc}, P^{cc})$. Interestingly, Home experiences higher welfare while social welfare under these policies than under the policies $(T^c, P^c)$. In a sense, the country with the binding self-enforcement constraint "blackmails" other countries by credibly threatening to slip out of the area of self-enforcing trade agreements.

The observation that terms-of-trade agreements may induce Home to move out of the zone of its voluntary compliance to the agreement constitutes the main finding of the present paper. The graphical analysis suggests that trade agreements targeting the world price are efficient only if the schedule defined by a binding self-enforcement constraint (locally) coincides with the individual indifference curves. In the next section, I will show that this latter assumption emerges from a naive extension of the standard repeated tariff game. It turns out, however, to be a knife-edge case.

3 The Model

In this section I will formalize the intuition derived from the diagrammatic analysis in the proceeding section. Specifically, I analyze under which conditions self-enforcing trade agreements may exclusively target the terms-of-trade without any loss of efficiency. These conditions are subsequently shown to be very restrictive.

3.1 The Setup

Consider a world economy consisting of two countries. Each country may set import tariffs $T$ and a generic domestic policy $P$. Nothing in the following analysis changes if $P$ represents an $n$-dimensional vector of policies. In the exposition, I will focus on the perspective of one country, which I will refer to as Home. All variables of the other country, Foreign, will be marked with an asterisk ($^*$).7

The countries engage in arm’s length trade of two goods, $X$ and $Y$. Home is assumed to be a net exporter of $X$ and a net importer of $Y$. The relative world price will be denoted

7With a small additional effort on notation, one may extend the following analysis to the case where Home trade with $N$ different countries.
by $\pi = p_X/p_Y$, where $p_X$ and $p_Y$ are the world prices of the two tradable goods net of import tariffs. The modeling setup is assumed to be entirely static.

Both countries maximize a generalized country-specific welfare function. It will be convenient to denote the flow of generalized welfare for Home and Foreign by

$$v(\pi, Z) \text{ and } v^*(\pi, Z^*).$$

where $\pi$ is the world price and the vectors $Z$ and $Z^*$ represent vectors of general domestic variables. The components of $Z$ could stand for local relative prices, real income, income distribution and other variables that can enter the political objectives of a government. While such a view is consistent with the current setup, I interpret $Z^*(*)$ as fundamental determinants of the model economy that pin down the production possibilities. Depending on the features of the underlying model, these determinants are the parameters of productivities, factor endowments and other determinants, which are sufficient statistics for all endogenous variables such as prices and income. I will return to this distinction below. For the time being, the variable $Z^*(*)$ is assumed to incorporate local policies and a vector of parameters $\xi^*(*)$, i.e.,

$$Z = (T, P, \xi) \text{ and } Z^* = (T^*, P^*, \xi^*)$$

The world price $\pi$ depends on all policies worldwide as well as on the economic determinants $Z(*)$, and can be written as $\pi(Z, Z^*)$.

I follow the literature by assuming that the only cross-border externality materializes through international prices. In particular,

$$\frac{d}{dT}v^*(\pi, Z^*)\bigg|_{\pi=\text{const}} = \frac{d}{dP}v^*(\pi, Z^*)\bigg|_{\pi=\text{const}} = 0$$

The corresponding equation holds for Home’s welfare. This setup allows for politically augmented local welfare functions and generalizes the setup in Bagwell and Staiger (2001).

Social welfare is an increasing function of foreign and domestic welfare. For simplicity and without loss of generality for what follows, I define the flow of social welfare as the

---

8I specifically do not exclude factors such as physical or human capital that are accumulated over time, but I assume that the stocks are in steady state and therefore do not bear time-subscripts.
following function of country welfare

\[ W(T, T^*, P, P^*) = [v(\pi, Z)]^\mu [v^*(\pi, Z^*)]^{1-\mu} \]  

(4)

with \( \pi = \pi(Z, Z^*) \) and \( \mu \in (0.1) \).

### 3.2 Trade Agreements

Having described the fairly general setup of the world economy, I next turn to the optimal design of trade agreements.

**The General Concept** Fundamental to the concept of trade agreements is the optimal tariff argument, according to which large countries are tempted to exert their market power on the world market by unilaterally charging tariffs, thereby distorting the terms-of-trade in their favor.\(^9\) This means that, starting from zero or low levels of tariffs, each country has an incentive to unilaterally raise tariffs. If all countries concede to these incentives, the resulting equilibrium is characterized by high tariffs, which inflict substantial efficiency losses on the world economy. This situation is often referred to as a tariff war.

I will not discuss the general features of trade wars here but simply introduce the notation \( T^N, T^*N, P^N \) and \( P^*N \) for the policies in the uncooperative equilibrium in which both countries set policies as in a one-shot Nash equilibrium. The resulting welfare functions in the trade war will be normalized to zero so that (4) grants that the surplus of cooperation is split in positive shares.

As countries trade repeatedly, trade agreements can remedy the inefficiencies of a tariff war in the spirit of repeated games. While aiming to reduce distortionary tariffs, they must account for countries’ defection temptations resulting from the optimal tariff argument. Specifically, a unilateral defection from a trade agreement takes the trade partner by surprise so that, for a limited transitional period, the defecting country enjoys the benefits of unilateral deviation.

After the transitory period of defection, however, a trade war ensues, the efficiency losses of which constitute the cost of defection. Self-enforcing trade agreements require that, for each country, the transitory gains from defection be exceeded by future losses resulting from the optimal tariff argument.

---

\(^9\)It is irrelevant if countries exploit their monopoly power through export taxes or their monopsony power through import tariffs (see Helpman and Krugman 1987).
from a tariff war. This incentive compatibility requirement is called the self-enforcement constraint.

The Self-Enforcement Constraint Consider a trade agreement that aims to implement the policies \((\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*)\), which result in the world price \(\bar{\pi}\) and the general long-run economic conditions \(\bar{Z}\). Foreign cannot adjust its policies for a period of positive length. However, after this transitory period, a trade war breaks out. Formally, the self-enforcement constraint for Home is

\[
\Gamma(\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*) = \sum_{k \geq 0} \beta^k \left[ v(\bar{\pi}, \bar{Z}) - v(\pi^{dk}, Z^{dk}) \right] \geq 0
\]  

where \(\pi^{dk} = \pi(Z^{dk}, Z^{*,dk}), Z^{(s),dk}\) are the economic conditions \(k\) periods after Home’s defection, involving \(T^{dk}, T^{*,dk}, P^{dk}\) and \(P^{*,dk}\), which represent Home’s uncooperative policies; \(\beta\) is Home’s discount factor. Inequality (5) requires that the discounted flow of domestic welfare, when respecting the trade agreement, weakly exceed the discounted flow of domestic welfare when defecting on it.

Introducing the notation \(T^{dk}\) and \(P^{dk}\) in addition to the Nash strategies \(T^N\) and \(P^N\) is not redundant for the following reason: foreign policies do not adjust instantaneously but after a time-lag. Therefore, for a transition period after defection, Home can set unilaterally optimal policies without suffering immediate retaliation. Hence, if Home defects in period 0, its unilaterally optimal policies in the transition period differ from the Nash policies.

The self-enforcement constraint for the foreign country is defined parallel to (5). In the following analysis, however, I assume that foreign policies do not suffer enforcement problems and neglect Foreign’s self-enforcement constraint.

Optimal Trade Agreements – The Traditional View The socially optimal trade agreement maximizes (4) subject to (5); it solves

\[
\max_{\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*} W(\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*) \quad \text{s.t.} \quad \Gamma(\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*) \geq 0
\]

10 Indeed, the reaction time of Foreign’s policies is the defining element of a period.

11 Writing down the global welfare function (4) and the constraint (5), I have implicitly limited the analysis to trade agreements with time-invariant policies, consistent with the current model’s static setup.

12 One may think of this case as \(\beta^* \to 1\), where \(\beta^*\) is Foreign’s discount factor.
For Home, the self-enforcement constraint is assumed to be binding. The optimal trade agreement then implements the policies satisfying

\[
\frac{W_X}{W_Y} = \frac{\Gamma_X}{\Gamma_Y} \quad X, Y \in \{\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*\}
\]  

(6)

In the following analysis, I assume that Foreign’s equilibrium policies \(\bar{T}^*\) and \(\bar{P}^*\) are implemented and sustainable. Under this assumption, I examine Home’s incentives and optimal policies.

The standard assumption of models in self-enforcing trade agreements is that, in the periods following defection, a trade war breaks out with all countries playing Nash strategies; moreover, the model parameters \(\xi^{(e)}\) from (2) are entirely exogenous and time-invariant. Under these assumptions, the function \(\Gamma\) from (5) becomes with (2)

\[
\Gamma(\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*) = (1 - \beta)^{-1} v(\pi, (\bar{T}, \bar{P}, \xi)) ... \\
\ldots - v(\pi^{d1}, (T^{d1}, P^{d1}, \xi)) ... \\
\ldots - \beta(1 - \beta)^{-1} v(\pi^N, (T^N, P^N, \xi))
\]  

(7)

The policies \(T^{d1}\) and \(P^{d1}\) are Home’s unilaterally optimal policies one period after defection, i.e., when Foreign maintains its cooperation policies \(\bar{T}^*\) and \(\bar{P}^*\). \(\pi^{d1}\) is the resulting world price.

When defining now the discounted flows of utility under cooperation and defection as

\[
V^C = \frac{\beta}{1 - \beta} v(\pi, (\bar{T}, \bar{P}, \xi)) \quad \text{and} \quad V^D = v(\pi^{d1}, (T^{d1}, P^{d1}, \xi)) + \frac{\beta}{1 - \beta} v(\pi^N, (T^N, P^N, \xi))
\]  

(8)

the self-enforcement condition (5) is simply \(V^C - V^D \geq 0\). Using these expressions, the ratio \(\Gamma_{\bar{P}}/\Gamma_{\bar{T}}\) is quickly computed so that the optimality condition (6) implies

\[
\frac{d}{dt} V^C + \frac{1 - \mu}{\mu} \frac{d}{dt} V^{*C} = \frac{d}{dt} V^C
\]  

(9)

With (4), this condition is

\[
\begin{bmatrix}
\frac{v_{\pi} + \frac{1 - \mu}{\mu} v_{\pi}^*}{v_{\pi} + \frac{1 - \mu}{\mu} v_{\pi}^*} \\
\frac{v_{\pi} + \frac{1 - \mu}{\mu} v_{\pi}^*}{v_{\pi} + \frac{1 - \mu}{\mu} v_{\pi}^*}
\end{bmatrix}
\begin{bmatrix}
\bar{\pi}_P + v_{\pi} Z_P \\
\bar{\pi}_T + v_{\pi} Z_T
\end{bmatrix}
\begin{bmatrix}
v_{\pi} \bar{\pi}_P + v_{\pi} Z_P \\
v_{\pi} \bar{\pi}_T + v_{\pi} Z_T
\end{bmatrix}
\]
or, after straightforward simplification,

\[ \frac{v_Z Z_P}{v_Z Z_T} = \frac{\bar{\pi}_P}{\bar{\pi}_T} \quad (10) \]

Notice that this optimality condition is written in a slightly cumbersome way and could be simplified with the expressions \( Z_T = (1, 0, 0, \ldots)^T \) and \( Z_P = (0, 1, 0, \ldots)^T \). For later use, however, it will be convenient to keep the current notation.

An important feature of defection policies is that they are independent of the cooperation policies \( \bar{T} \) and \( \bar{P} \), as is the entire welfare under defection (the second and third term on the right hand side of (7)). Therefore, just as illustrated in Figure 4, the schedule of Home’s policies for which the self-enforcement constraint holds with equality coincides with its indifference curve. The formal analysis thus shows that the case illustrated in Figure 4 is not purely random but reflects the assumptions of the standard framework that the literature has used to analyze self-enforcing trade agreements.

In sum, Home’s policies that marginally satisfy the self-enforcement constraint coincides with its indifference curve. This feature has consequences for the specific type of trade agreement considered next.

**Terms-of-Trade Agreements – the Traditional Framework.** Now consider a trade agreement that only requires countries to choose policies that are compatible with a predetermined level of terms-of-trade \( \bar{\pi} \) while leaving the exact policy mix to the individual countries. A natural question then emerges: is such a trade agreement targeting the terms-of-trade \( \bar{\pi} \) from above enforceable? To answer this question, consider Home’s welfare along the schedule defined by \( \pi = \bar{\pi} \). Given that Home respects this trade agreement, it will choose the policy mix that solves the program (2)

\[
\max_{\bar{T}, \bar{P}} v(\pi, (\bar{T}, \bar{P}, \xi)) \quad s.t. \; \pi = \bar{\pi}
\]

The corresponding first-order conditions are, when written slightly unconventionally way with the help of the definition of \( Z \) from (2),

\[ \frac{v_Z Z_P}{v_Z Z_T} = \frac{\bar{\pi}_P}{\bar{\pi}_T} \quad (11) \]
The condition of Home's individual optimum under cooperation (11) is thus identical to the optimality condition of the social optimum (10). Thus, Home's optimal strategy, conditional on respecting the trade agreement, coincides with that of the social optimum. Moreover, the defection value is independent of the cooperation policies \((\bar{T}, \bar{P})\) so that Home's self-enforcement constraint is satisfied at the social optimum. Therefore, Home is indifferent to respecting or defecting on the trade agreement that fixes the terms-of-trade to \(\bar{\pi}\). Overall, the observations imply that the trade agreement that only targets the terms-of-trade \(\pi\) is self-enforcing.

In sum, and just as shown in Figure 4, the optimal trade agreement within the class of trade agreements fixing the terms-of-trade coincides with the optimal trade agreement. There is no efficiency loss when negotiating the terms-of-trade instead of targeting the precise policy mix.

This result confirms and generalizes the findings of Bagwell and Staiger (2001), who argue that granting countries the flexibility to freely choose their policy mix while preserving a given level of terms-of-trade "would ensure that governments could achieve efficient trade and domestic policy outcomes." This result, however, depends on one crucial assumption, which I will discuss next.

### 3.3 Persistent Economic Conditions

All results established thus far have been derived under the implicit assumption that the variables \(Z^{(e)}\) from (2) depend only on current domestic policies. Recalling that the variables \(Z^{(e)}\) are supposed to reflect deep economic determinants such as productivities, factor endowments and physical or human capital stocks (possibly sector-specific), this assumption is questionable. While this assumption is common across the literature,\(^{13}\) it is only compatible with an economy that is rebuilt or reorganized from scratch in every period. Following the more natural view, one should assume that capital stocks depend on past investment decisions and thus on past policies. This assumption holds for human as well as for physical capital stocks, for the overall capital stocks as well as sector-specific capital stocks, which may be irreversible. Moreover, labor allocation may be sluggish to adjust.

---

\(^{13}\)Starting from Dixit (1987), virtually all work on self-enforcing trade agreements has built on this assumption.
meaning that the current occupational composition of an economy depends on past labor market policies. In sum, past economic decisions and policies are bound to determine current production possibilities and the corresponding economic outcomes. These economic outcomes include the optimal policies as well as the payoff in the case of defection. Contemplating these intertemporal linkages that affect an economy, the view that the deep economic variables \( Z(t) \) defined in (2) is not a realistic view of the world.

The assumption that economic conditions are independent of past policies is certainly unrealistic. Unfortunately, it also turns out to be crucial for the assessment of self-enforcing trade agreements. The logic is as follows: past policies affect past economic conditions and past economic conditions, in turn, determine current production possibilities. The production possibilities finally shape the defection payoffs and therefore the self-enforcement constraints. In this case, the condition (10) no longer describes the social optimum and the consequences for the efficiency of terms-of-trade agreements are severe.

To formally capture the intertemporal linkages of the economies in a common way, I treat the variable \( Z(t) \) as a regular state variable that depends not only on present domestic policy variables but also on the past realizations of \( Z(t) \). Specifically, I adapt the definitions of \( Z \) in (2) for a given \( t \) adapted to

\[
\xi_t = h(Z_{t-1})
\]

and the parallel for \( Z^* \). With the thus altered state variable \( Z \), I denote with \( \bar{Z} \) the steady-state value of \( Z \) given the cooperation strategies \( \bar{T}, \bar{T}^*, \bar{P} \) and \( \bar{P}^* \). Furthermore, let the corresponding steady state of \( \xi \) be denoted by \( \bar{\xi} \) (\( \bar{\xi} = h(\bar{T}, \bar{P}, \bar{\xi}) \)). The self-enforcement constraint then becomes

\[
\Gamma(\bar{T}, \bar{T}^*, \bar{P}, \bar{P}^*) = (1 - \beta)^{-1} v(\bar{\pi}, (\bar{T}, \bar{P}, \bar{\xi}))... \\
... - \sum_{k \geq 0} \beta^k v(\pi^{dk}, (T^{dk}, P^{dk}, h(Z_{k-1})))
\]

In this expression \( T^{dk} \) and \( P^{dk} \) stand for Home’s unilaterally optimal policies in period \( k \geq 0 \), while \( \pi^{dk} \) stands for the resulting world price. In period \( k = 0 \), Foreign is still

---

14 Of course, expected future policies affect \( Z \) as well.
15 If there is any need to underpin this last statement, see Kennan and Riezman (1988), Devereux (1997) and Sauré (2011) and (2012).
assumed to stick to $\bar{T}^*$ and $\bar{P}^*$, but in the consecutive periods, it plays uncooperatively. The intertemporal linkages imply that $Z_t$ depends on the time elapsed since defection, which in turn means that the uncooperative strategies must now bear a time subscript.

The discounted flows of utility under cooperation and defection are, similar to (8)

$$V^C = (1 - \beta)^{-1} v(\bar{\pi}, (\bar{T}, \bar{P}, \bar{\xi})) \quad \text{and} \quad V^D = \sum_{k \geq 0} \beta^k v(\pi^{dk}, (T^{dk}, P^{dk}, h(Z_{k-1})))$$

However, Home’s flow of utility under defection, $V^D$, now depends on past realizations of $Z$ and through the dependence of $Z$ on past policies. In particular, for $X = \bar{T}, \bar{P}$ the derivative

$$\frac{d}{dX} V^D = \sum_{k \geq 0} \beta^k \frac{d}{dX} v(\pi^{dk}, h(T^{dk}, P^{dk}, Z_{k-1})) = \sum_{k \geq 0} \beta^k \left[ v_Z \frac{d\pi^{dk}}{dZ_k} + v_Z \right] h_Z \frac{dZ_{k-1}}{dX}$$

is different from zero as long as neither $h_Z = dh(T, P, Z')/dZ'$ nor the derivatives $dZ_{k-1}/d\bar{X}$ are zero for all $k \geq 0$. Quite intuitively, expression (13) shows that whenever past policies affect the economic decisions that determine future production possibilities, those policies shape defection utilities.

Computing now the ratio $\Gamma_{\bar{P}}/\Gamma_{\bar{T}}$ to establish (6) with (4) thus renders

$$\frac{d}{dT} V^C + \frac{1 - \mu}{\mu} \frac{d}{dT} V^{*C} = \frac{d}{dT} V^C - \frac{d}{dT} V^D$$

In contrast to (9), Home’s defection utility, $V^D$, now enters the condition for the social optimum in economies with intertemporal linkages, (14).

When respecting the terms-of-trade agreement implementing $\bar{\pi}$, Home’s individual welfare is maximized at

$$\max_{T,P} v(\pi, (\bar{T}, \bar{P}, \bar{\xi}(\bar{T}, \bar{P}))) \quad \text{s.t.} \quad \pi = \bar{\pi}$$

The corresponding first-order conditions are again (11).

A straightforward computation confirms the following equivalence:

$$\frac{d}{dT} V^C = \frac{d}{dT} V^D \quad \Leftrightarrow \quad \frac{v_Z Z_P}{v_Z Z_T} = \frac{\bar{\pi}_P}{\bar{\pi}_T}.$$  

The optimum of the relative impact of Home’s policies on its defection is exactly the same as their relative impact on its cooperation utility. Is this condition naturally satisfied? In fact, it is not.
To take a very simple example, consider an economy where production technologies of the two goods, \( X \) and \( Y \), exhibit the same unit factor requirements. International trade and specialization, which may be driven by Ricardian differences in factor productivities, does not affect the interest rate. Therefore, the steady-state capital stock of an economy is not affected by the degree of trade integration and thus by the policy \( T \). Now consider the effect of the domestic policy \( P \), representing an investment subsidy. Such a subsidy will increase the steady-state level of the capital stock and, by affecting income, alter Home’s marginal welfare (or utility) under defection. In this case, the left hand side of (15) is certainly violated because \( dV^D/dT = 0 \) but \( dV^D/dP \neq 0 \). This example can easily be generalized to investment subsidies for sector-specific factor accumulation, which affects the industrial mix of a country in different ways than a simple tariff reduction would with different effects of defection welfare, to labor market policies when labor allocation is partly rigid.\(^{16}\)

In general, because policies with intertemporal effects that are optimal under cooperation have entirely different marginal effects under defection, the left-hand side of (15) is almost surely violated. In this sense, the socially optimal and Home’s individually optimal policy mix coincide only in knife-edge cases. One particular example of those cases is the situation where \( h \) from (12) is the constant function, i.e., where economic conditions are not intertemporally linked; this is the case on which existing studies of self-enforcing trade agreements have focused.

**Terms-of-Trade Agreements.** Now consider the type of trade agreements that target the terms-of-trade \( \pi \), leaving the maximization of individual welfare to each country. In particular, consider the trade agreement requiring Home to choose a mix of policies that does not alter the terms-of-trade \( \pi^c \). Otherwise, Home is allowed to freely choose its policies. This liberty will result in Home maximizing its welfare, which implies that (11) holds. I label this policy mix \((T^d, P^d)\) (in line with Figure 5). By the definition of individual maximization, this strategy renders strictly higher welfare for Home. Moreover, Foreign welfare is unchanged as long as \( \pi^c \) is constant. Therefore, social welfare, conditional on cooperation, increases as Home’s policies move from \((T^c, P^c)\) to \((T^d, P^d)\). Consequently this

\(^{16}\)See also the discussion of strategic dependence in Sauré (2011) or Sauré (2012) for an elaborate example where production subsidies can undermine the viability of a terms-of-trade agreement.
move increases the social welfare, which by the very definition of \((T^c, P^c)\) as the constrained optimum, implies that \((T^d, P^d)\) cannot be self-enforcing. In sum, the social optimum cannot be implemented by trade agreements that target terms-of-trade only. These observations imply the following proposition.

**Proposition 1** If a self-enforcement constraint binds, terms-of-trade agreements are almost surely inefficient.

The proposition formulates the present paper’s central result. The following intuition should be quite clear by now: within the confines of an established terms-of-trade agreement, Home’s policy mix generally affects its defection value. Thus, by optimizing its policy mix within the boundaries of such an agreement, Home may slip out of the zone where it voluntarily complies with the agreement and consequently defect (compare Figure 5). This possibility implies that the set of self-enforceable terms-of-trade agreements constitute a subset of the larger set of self-enforceable trade agreements that target policies directly. The optimum within the subset results in efficiency losses vis-à-vis the general optimum.

The existing literature on efficient trade agreements has consistently reported results standing in clear opposition to Proposition 1. Specifically, Bagwell and Staiger (2001) show that, "conditional on a level of market access, there is simply no reason for a government to distort – in light of its own objectives – to policy mix with which it delivers that market access, and this policy mix is irrelevant to the objectives of its trading partner." The authors argue that granting countries the flexibility to freely choose their policy mixes while preserving a given level of terms-of-trade "would ensure that governments could achieve efficient trade and domestic policy outcomes." Further, Ederington (2001) states that "any change in the temptation to deviate is most efficiently handled by adjusting trade policy" while domestic policies are set to their individually optimal levels.

What drives these diametrically opposite findings? Existing studies have failed to recognize the importance of intertemporal linkages. While these linkages are the bread and butter of traditional macroeconomic literature, the work on self-enforcing trade agreements has neglected them. Instead, it has built on the well-known folk theorem that states that the inefficiencies of the prisoner’s dilemma can be remedied in a repeated game. The trade literature has not only adopted the spirit of repeated games but has also stuck to its textbook
approach in the sense that the same game with the same payoff conditions is played over and over again. In particular, the literature of repeated trade-games has held to models where current strategies impact current payoffs only. This approach entails the implicit assumption that in each period, economic agents act within an entirely new, re-organized economy, without any "history" in the basic macroeconomic sense. While this assumption may be an admissible approximation to describe the general principles of cooperation, it is ill-suited for analyzing the efficient design of trade agreements – particularly those involving more than one policy.

One additional comment on Proposition 1 is in order. Throughout my analysis, I have focussed on Home’s self-enforcement constraint and Home’s options to slip along the iso-price schedule ($\pi^c = \text{const.}$), thereby moving out of the region where its own self-enforcement constraints are binding. I should stress that it is not a coincidence that the self-enforcement constraint and the defection temptations of one and the same country are analyzed. In fact, Home will ignore its own self-enforcement constraint and defect whenever it is individually beneficial; Home does not see its own constraint. The case is different for the self-enforcement constraints of foreigners. Home will be eager not to choose policies that induce a violation of foreign self-enforcement constraints because foreign defection would clearly inflict losses for Home. Therefore, it is in Home’s individual interest to make foreign countries comply with the trade agreement. Home will never let its policies slip into a region where the self-enforcement constraints of foreigners are violated.

Finally, one may wonder which characteristics define the optimal trade agreement and how complicated this agreement is to achieve. After all, a striking virtue of terms-of-trade agreements is their simplicity. The answer is, of course, encapsulated in the optimality condition (14). While the traditional approach postulates that the only important consideration concerns the policies’ marginal impact on Home’s cooperation welfare (compare (10)), condition (14) shows instead that the marginal impact of the policies on the welfare differential between cooperation and defection. Specifically, the optimal trade agreement equalizes the relative marginal impact of policies on the individual welfare differential with their relative marginal impact on the social optimum. Clearly, this rule lacks the convenient simplicity of the terms-of-trade agreements, but the resulting efficiency gains may be worth the effort.
4 Conclusion

The existing literature on trade agreements has argued that efficient trade agreements only target terms-of-trade and leave domestic policies to the free choice of national governments. That claim builds on the crucial assumption that current economic conditions are independent of past policies. This assumption blatantly contradicts the common macroeconomic understanding that key economic factors such as capital stocks, skill levels or labor allocation are slow to adjust and, at the same time, subject to policy intervention. Allowing instead for some degree of persistence in economic conditions defies the efficiency of terms-of-trade agreements. Persistent effects of policies interact with the temptation to defect on trade agreements and thus must be targeted in efficient trade agreements. Therefore, efficient trade agreements must target the mix of domestic policies directly as soon as the latter affect a country’s defection temptations. This is the case with the exception of knife-edge cases.

Some of the common knowledge remains of course valid. For example, the finding that international trade agreements do not induce a "race to the bottom" of labor standards stays intact (see Bagwell and Staiger (2001)) as long as such dynamics do not interfere with enforcement problems. I do show, instead, that enforcement issues constitute grounds to target the exact mix of domestic policies in the legal text of efficient trade agreements. Therefore, it seems premature to conclude with Bagwell and Staiger (2006) that WTO rules, which target production subsidies, "may ultimately do more harm than good to the multilateral trading system."

Reading the present paper’s results in a general and pragmatic way, one would not want to include all possible domestic policy instruments in the coming WTO negotiation rounds. However, one may need to incorporate those domestic policies that seriously influence defection temptation and that may therefore undermine the willingness to comply with international treaties. Ignoring these policies comes at efficiency losses.
References


Sauré, P. 2011. How to Use Subsidies to Sustain Trade Agreements. SNB WP

Sauré, P. 2012. On Subsidies in Trade Agreements. SNB mimeo