Regulation via the Polluter-Pays Principle*

Stefan Ambec† Lars Ehlers‡

December 2010

Abstract

We consider the problem of regulating an economy with environmental pollution. We examine the distributional impact of the polluter-pays principle which requires that any agent compensates all other agents for the damages caused by his or her (pollution) emissions. With constant marginal damages we show that regulation via the polluter-pays principle leads to the unique welfare distribution that assigns non-negative individual welfare and renders each agent responsible for his or her pollution impact. We extend both the polluter-pays principle and this result to increasing marginal damages due to pollution. We also discuss the acceptability of the polluter-pays principle and compare it with the Vickrey-Clark-Groves mechanism.

JEL classification: C7, D02, D30, D6.

Keywords: Regulation, Polluter-Pays Principle, Responsibility for Pollution Impact, Externalities.

*This research started while the second author was visiting the Toulouse School of Economics (INRA-LERNA). It received financial support from INRA (France), the ANR (France) through the project ANR-08-JCJC-0111-01 on "Fair Environmental Policies", the SSHRC (Canada) and the FQRSC (Québec).

†Toulouse School of Economics (INRA-LERNA-IDEI), France, and University of Gothenburg, Sweden; e-mail: stefan.ambec@toulouse.inra.fr.

‡Département de Sciences Économiques and CIREQ, Université de Montréal, Canada; e-mail: lars.ehlers@umontreal.ca (Corresponding author).

1 Introduction

From water management to air pollution, managing environmental problems efficiently requires well-designed public policies or coordination among stakeholders (Ostrom, 1990). Environmental policies are launched to mitigate the failure of market economy due to the presence of negative externalities. Yet public intervention has an impact not only on the welfare of the economy as a whole but also on the distribution of welfare. This paper addresses the distributional impact of environmental policies in economies with pollution. The model allows for a variety of negative externalities including unilateral or multilateral ones, heterogenous impacts due to distance or mitigation. It formalizes many complex environmental issues such as water quality management in a river or the reduction of sulfur dioxide or greenhouse gas emissions in an international setting.[1]

In this framework, we define a regulation mechanism as individual transfers contingent on pollution emissions. In particular, we consider the mechanism inspired by a literal interpretation of the polluter-pays (PP) principle. It states that *the costs of pollution should be borne by the entity which profits from the process that causes pollution.* Strictly speaking, it requires that any agent (firm or consumer) compensates all agents who suffer from his pollution emissions for the damage he causes. The PP mechanism is by construction budget-balanced. It is also efficient in the sense that it uniquely implements the allocation of pollution emissions that maximizes total welfare in Nash equilibrium. Therefore, the PP mechanism shares this feature with the mechanisms proposed by Duggan and Roberts (2002) and Montero (2008).

Note however that here each agent only chooses his emissions whereas in Duggan and Roberts (2002) each agent chooses his emission and reports the emission of his neighbor and in Montero (2008) each agent reports his inverse demand for any level of emissions. The focus of both papers is on the implementation of the efficient allocation under asymmetric information whereas we are interested in the distributional impacts of a mechanism that implements the efficient allocation under perfect information.

We examine the properties of the welfare distribution induced by regulation mechanisms. We focus on two fairness criteria. The first one is that individual's welfare is non-negative. It

[1] To that respect, it is as rich as the seminal model of Montgomery (1972).

is a minimal acceptability requirement since an agent who obtains a negative welfare does not benefit from the welfare-enhancing economy activities exhibiting pollution. The second criteria relies on the concept of responsibility in axiomatic theory of justice (Fleurbaey, 2008). It makes a polluter responsible for its pollution impact. More precisely, the welfare distribution should be such that a polluter is assigned the full social cost due to his pollution. In particular, if a polluter modifies the environmental impact of his own emissions in the economy, he should get the full return or loss due to this change. For instance, a firm which filters its own emissions to reduce their sulfur content should get the full benefit for the economy of its cleaning investment. A farmer who uses more pesticide and fertilizers leading to dirtier waste water should pay the social cost associated to this pesticide and fertilizer increase. We show that the welfare distribution induced by applying the PP mechanism is the only one that satisfies the two above criteria: non-negativity and responsibility for pollution impact.

We are also concerned by the acceptability of the PP principle. This is an important issue since environmental policies emerge as a collective choice in democratic societies. Similarly, international environmental agreements such as the Kyoto protocol are negotiated by sovereign countries. Each country is free to refuse any agreement that is worse than the status-quo or to join another agreement. We define acceptability as follows: a regulation mechanism is acceptable if no other regulation mechanism assigns to any agent or group of agents a strictly higher welfare (including no regulation at all). The PP mechanism fails to be acceptable in general. For instance, the most upstream polluter of a river prefers laissez-faire to the application of the PP principle. We nevertheless show that it is acceptable if externalities are multilateral among homogeneous agents. This is for instance the case in most theoretical models examining international agreements for greenhouse gas emission reduction including Chandler and Tulkens (1992), Carraro and Siniscalco (1993) and Barrett (1994).

Our first characterization of the polluter-pays welfare distribution relies on the assumption that marginal damage due to pollution does not depend on pollution concentration. When marginal damage is increasing with pollution concentration, the incremental impact of each polluter on damage is not straightforwardly defined. We extend the polluter-pays principle to this framework by making the polluter pay for the incremental impact of his or her emissions on society while victims of pollution are fully compensated for the damage caused by others.

We then provide a further characterization of the PP welfare distribution in the more general framework where marginal damage is increasing with pollution. More precisely, we show that it is the unique welfare distribution that satisfies three criteria: non-negativity, responsibility for pollution impact and solidarity upper bounds. The latter bounds require no agent to obtain more than what he or she would get in absence of pollution of all other agents. We conclude the paper by comparing the polluter-pays regulation mechanism with the Vickrey-Clark-Groves (VCG) mechanism applied to the pollution problem.

We proceed as follows. Section 2 introduces a simple model of pollution with constant marginal damages. It also provides several real-world examples which fit our framework. Section 3 describes regulation mechanisms and their induced distribution rules in equilibrium. It also discusses several regulations used in real life. Section 4 introduces the polluter-pays regulation and characterizes its induced distribution rule in terms of non-negativity and responsibility for pollution impact. Section 5 discusses the acceptability of the polluter-pays principle in societies. Section 6 generalizes our model to differentiate pollution and damages and allows for increasing marginal damages. We generalize the PP principle to this environment and extend our main results to this framework. Section 7 concludes by comparing the PP mechanism with the VCG mechanism.

2 A model of pollution

Consider a set $N = \{1, \ldots, n\}$ of agents (countries, cities, farmers, firms, consumers,...). Each agent $i \in N$ is polluting or is polluted or both. Agent i enjoys a benefit $b_i(e_i)$ from production and/or consumption where $e_i \geq 0$ denotes the level of economic activity hereafter called "emissions". The benefit function b_i is assumed to be both strictly concave and strictly increasing from 0 to a maximum $\hat{e}_i$ with $b_i'(\hat{e}_i) = 0$ for every $i \in N$,[2] and twice continuously differentiable: for all $i \in N$ and for all $0 \leq e_i < \hat{e}_i$, both $b_i'(e_i) > 0$ and $b_i''(e_i) < 0$. We normalize $b_i(0) = 0$ and assume that the marginal benefit at $e_i = 0$ is high enough (say infinite) so it is optimal for all agents to produce and/or to consume.

Pollution from agent i causes a marginal damage $a_{ij} \geq 0$ to agent j. The parameter

[2]This is without loss of generality since the maximum could be $\hat{e}_i = +\infty$ for some $i \in N$.

a_{ij} measures the magnitude of the pollution impact of i's emission on j. For the moment we consider constant marginal damages. Later we extend our results to environments with convex damages and thus, increasing marginal damage from emissions. A (negative) externality or pollution problem (N, b, a) is defined by a set of agents N, a profile of benefit functions $b = (b_i)_{i\in N}$, and a matrix of externality/pollution marginal impacts $a = [a_{ij}]_{ij\in N\times N}$. When there is no confusion, we write for short a instead of (N, b, a).

Let $Ri = \{j \in N | a_{ij} > 0\}$ denote the receptors of i's pollution: the set of agents which are polluted by i. Let $R^0i = \{j \in N\backslash\{i\} | a_{ij} > 0\}$ denote the receptors of i's pollution excluding i. We assume that $a_{ii} > 0$ for any $i \in N$ with $R^0i \neq \emptyset$, i.e. if i is polluting other agents, then his pollution also causes some damage at his location.[3] Let $Si = \{j \in N | a_{ji} > 0\}$ denote the set of agents who pollute agent i. Let $S^0i = \{j \in N\backslash\{i\} | a_{ji} > 0\}$ denote the set of agents who pollute i excluding i. The environmental damage suffered by i in the emission vector $e = (e_i)_{i\in N}$ is therefore

$$d_i = \sum_{j\in Si} a_{ji}e_j.$$

The welfare of agent i with emissions $e = (e_i)_{i\in N}$ is:

$$b_i(e_i) - d_i = b_i(e_i) - \sum_{j\in Si} a_{ji}e_j. \tag{1}$$

The first term in (1) is i's benefit from his own emissions whereas the second term is i's welfare loss due to pollution.

An *efficient* emissions plan $e^* = (e_i^*)_{i\in N}$ maximizes total welfare $\sum_{i\in N}[b_i(e_i) - d_i] = \sum_{i\in N} b_i(e_i) - \sum_{i\in N}\sum_{j\in Si} a_{ji}e_i$. It satisfies the following first-order conditions for every $i \in N$:

$$b_i'(e_i^*) = \sum_{j\in Ri} a_{ij}. \tag{2}$$

Note that our assumptions on the benefit function b_i guarantee that e_i^* is unique because $b_i'(\hat{e}_i) = 0$ and b_i is strictly concave and strictly increasing between 0 and $\hat{e}_i$. The marginal benefit of pollution emitted by i should be equal to its marginal damage for society. Let

$$W(a) = \sum_{i\in N} b_i(e_i^*) - \sum_{i\in N}\sum_{j\in Si} a_{ji}e_i^*$$

[3]For the environments considered here, this assumption is without loss of generality.

denote the economy's welfare from the efficient emissions plan e^* in the problem (N, b, a). A welfare distribution for the problem (N, b, a) is a vector $z = (z_i)_{i \in N}$ such that $\sum_{i \in N} z_i \leq W(a)$. A distribution rule ϕ associates with any problem (N, b, a) a welfare distribution $\phi(a)$ for a. Note that a distribution rule identifies for each problem a welfare distribution which the society may wish to implement.

The externality problem (N, b, a) exhibits multilateral externalities if $Si = Ri$ for any $i \in N$. The problem (N, b, a) exhibits unilateral externalities if $S^0 i \cap R^0 i = \emptyset$ for any $i \in N$. Let $V \subseteq N$ denote the set of agents who do not pollute other agents and only suffer from pollution due to other agents' activities. Formally, for any $i \in V$, $a_{ij} = 0$ for all $j \neq i$ and $a_{ji} > 0$ for at least one $j \neq i$, or equivalently $R^0 i = \emptyset$ and $S^0 i \neq \emptyset$; and without loss of generality, $\hat{e}_i = 0$.[4] Similarly, let $P \subseteq N$ denote the set of agents who do not suffer from other agents' pollution: $a_{ij} > 0$ for some $j \neq i$ and $a_{ji} = 0$ for all $j \neq i$, that is $R^0 i \neq \emptyset$ and $S^0 i = \emptyset$. Note that any agent in $N \backslash V$ is polluting the society from his economic activities.

Example 1 (The River Pollution Problem) Agents are countries, cities or factories located along a river. The set of predecessors of i in the river is $S^0 i$ while the set of followers of i is $R^0 i$. Each agent i emits e_i units of pollution which impact its followers downstream: one unit emitted in i causes a marginal damage a_{ij} in j. Here the marginal damages a_{ij} may be decreasing with respect of the distance of j to i, i.e. agent i's emissions have a higher pollution impact on immediate neighbors than on agents located further downstream the river. Symmetrically, agent i suffers from pollution emitted upstream by agents in $S^0 i$ and by himself.[5] It is a case of unilateral externalities: if we take two agents i and j, either i is upstream j or i is downstream j, i.e. $i \in Rj$ or $i \in Sj$. In a single canal or one-tributary river, agents can be ordered according to their position from upstream to downstream. In this case, if $N = \{1, ..., n\}$ and if agents suffer from their own pollution (e.g. countries), then for any $i \in N$, $Ri = \{1, 2, \ldots, i\}$ and $Si = \{i, i+1, \ldots, n\}$. Moreover, for any i and j, if $j \in Ri$ then $Rj \subseteq Ri$. Symmetrically, if $j \in Si$, then $Sj \subseteq Si$. The latter properties might not hold in

[4] If $\hat{e}_i > 0$, then agent i's activity does not have any impact on society and his activities can be disregarded.

[5] In the case of a river, "linearity is a good approximation up to the point at which the river becomes so overloaded with organic material that oxygen (needed for aerobic bacteriological decomposition) is depleted. At that point, [refereed as the river carrying capacity] the river's capacity to clean itself is greatly diminished." from Kolstad (2000) footnote 2 page 177.

more general rivers. With several tributaries that end up on the same main course, for any agent i there might be $k, j \in Si$ but both $k \notin Sj$ and $j \notin Sk$. Symmetrically, for river deltas or irrigations ditches originated from a source or weed or reservoir, we have the reverse: for any agent i there might be $k, j \in Ri$ but $k \notin Rj$ and $j \notin Rk$.[6]

Example 2 (The International Greenhouse Gas Emissions Game) Players are countries. Each country i enjoys a benefit b_i from its own greenhouse gas emissions e_i. Greenhouse gases emitted into the atmosphere cause global warming that damages countries' economies. The magnitude of global warming depends on total emissions on the earth surface $\sum_{j \in N} e_j$. Suppose that total emissions cause a constant marginal damage of δ_i to country i. In this example, $Si = Ri = N$ and $a_{ii} = a_{ij} = \delta_i$ for all $i, j \in N$: all countries exert multilateral externalities on all other countries of the same magnitude. Yet countries differ on the damage that externalities cause on their economy. Seminal papers on international agreements for greenhouse emission reduction (Chandler and Tulkens, 1992; Carraro and Siniscalco, 1993; Barrett, 1994) rely on these assumptions except that they consider convex damage (or concave benefit of emission abatements) and, therefore, increasing marginal damage.

Example 3 (The International Acid Rain Game) Agents are countries emitting sulfur dioxide (SO2) by burning coal for power production. This causes acid rain which damages forests and ecosystems in neighboring countries. The parameter a_{ij} captures the marginal impact of country i's SO2 emissions to acid rain in country j. It depends on the fraction of emissions from i that is deposited in j and its marginal damage on j. Mäler and De Zeeuw (1998) provide estimations on those parameters for 1990 and 1991 in Europe. For instance, among the SO2 emissions from Belgium, 19.4% ended up in Belgium, 13.3% in Germany, 9% in France, 4.8% in Netherland and so on. Mäler (1989, 1994) considers an acid rain game with such heterogeneous "transportation" parameters and constant marginal damage. This game has been extended by Mäler and De Zeeuw (1998) and Finus and Tjøtta (2003) to environments with convex marginal damages.

Example 4 (Polluters versus Victims) Agents in V are individuals and those in P are

[6] See Ambec and Sprumont (2002) and Ambec and Ehlers (2008) for a rigorous analysis of the river water sharing problem.

firms and each agent belongs either to V or to P. Firms emit pollution without incurring any damage: $a_{ij} = 0$ for every $j \neq i$. In contrast, any $i \in V$ does not emit pollution but suffers from pollution: $\hat{e}_i = 0$ for every $i \in V$ and $a_{ji} > 0$ for at least one $j \in P$. In this case, a_{ji} can be interpreted the marginal damage of each unit of firm j's pollution causes to person i in term of health or environmental impact. It depends on technologies, distance between firms and individuals, climatic conditions, and so on. The victims of pollution might be firms involved in different sectors than the polluter ones; for instance hotel and restaurants located close to a lake or sea shore that might be polluted by local factories. The main difference with the previous examples is that emitters and victims are disjunct sets of agents. It is a case of unilateral externalities.

3 Regulation mechanisms and distribution rules

An important policy tool in pollution problems are regulation mechanisms. A regulation mechanism $t : \mathbb{R}^N_+ \longrightarrow \mathbb{R}^N$ specifies for any emissions a vector of payments (or transfers) $t(e) = (t_i(e))_{i \in N}$. It assigns to agent i the transfer $t_i(e)$ for any emissions plan $e = (e_i)_{i \in N}$. Given the mechanism t and the emission plan e, agent i's welfare under the vector $t(e)$ is given by

$$b_i(e_i) - d_i + t_i(e) = b_i(e_i) - \sum_{j \in Si} a_{ji} e_j + t_i(e). \tag{3}$$

Of course, each agent i chooses his own emissions and for any problem a, the regulation mechanism t induces an "emissions revelation game". Let $\mathcal{N}(t, a)$ denote the set of (pure) non-cooperative Nash equilibria in the emissions game under the mechanism t and the problem a.

In the non-cooperative Nash equilibrium of the externality problem with the mechanism t, each player i maximizes (3) with respect to e_i given $e_{-i} = (e_j)_{j \in N \setminus \{i\}}$. Let $e^t \in \mathcal{N}(t, a)$ be a Nash equilibrium emission plan. Agent i's equilibrium welfare under e^t is:

$$z_i^t = b_i(e_i^t) - d_i^t + t_i(e^t),$$

where $d_i^t = \sum_{j \in Si} a_{ji} e_j^t$. The total welfare is

$$W^t(a) = \sum_{i \in N} z_i^t = \sum_{i \in N} \left[b_i(e_i^t) - d_i^t + t_i(e^t) \right] = \sum_{i \in N} b_i(e_i^t) - \sum_{i \in N} d_i^t + \sum_{i \in N} t_i(e^t),$$

where in the last expression the first term is the total benefit from emission, the second is the total damage and the third is the regulation mechanism surplus (or deficit if negative).

Given a distribution rule ϕ and a mechanism t, we say that t implements ϕ (in Nash equilibrium) if for all problems a and all $e^t \in \mathcal{N}(t,a)$, we have

$$\phi_i(a) = z_i^t = b_i(e_i^t) - \sum_{j \in Si} a_{ji} e_j^t + t_i(e^t).$$

A particular regulation mechanism is the *laissez-faire* mechanism t^{lf} defined by $t_i^{lf}(e) = 0$ for all $i \in N$ and all $e \in \mathbb{R}_+^N$. The laissez-faire mechanism represents situations without regulation or where society chooses not to intervene. It implements the emissions plan $e^{lf} = (e_i^{lf})_{i \in N}$ satisfying the following first-order conditions,

$$b_i'(e_i^{lf}) = a_{ii},$$

for every $i \in N$. Thus, for each problem a, $\mathcal{N}(t^{lf}, a)$ is unique and implicitly given by the above equalities. In contrast to the efficient emissions plan e^*, under laissez-faire each agent i considers the impact of his emissions only on his own welfare. In particular, $e_i^{lf} = \hat{e}_i$ if $a_{ii} = 0$. As long as $a_{ij} > 0$ for some $j \neq i$, i.e. i's emissions have an impact on another agent j, then $e_i^{lf} > e_i^*$ and therefore $d_j^{lf} > d_j^*$ for every $j \in Ri$.

Many regulation mechanisms are used in practice. For instance, consider a norm on pollution emissions mechanism, denoted by $\bar{t}$. It defines upper bounds on emissions $\bar{e}_i \geq 0$ and penalties for exceeding these bounds. Formally, let $\bar{e} = (\bar{e}_i)_{i \in N}$ and for all $e \in \mathbb{R}_+^N$,

$$\bar{t}_i(e) = \begin{cases} 0 & \text{if} \quad e_i \leq \bar{e}_i \\ -F_i(e_i - \bar{e}_i) & \text{if} \quad e_i > \bar{e}_i \end{cases}$$

for every $i \in N$ where $F_i > 0$ is the fine in case of excess pollution (which can be infinite or lump-sum). In case of an uniform norm, $\bar{e}_i = \bar{e}_j$ and $F_i = F_j$ for all $i, j \in N$. If the fine is high enough to be persuasive and the norm is binding in the sense that $e_i^{lf} > \bar{e}_i$ for all $i \in N$, then the unique emissions plan implemented in Nash equilibrium by $\bar{t}$ are $e_i^{\bar{t}} = \bar{e}_i$ for all $i \in N$.

The emission fee mechanism t^f specifies fees $f = (f_i)_{i \in N}$ on emissions and, therefore, charges the payment $t_i^f(e) = -f_i e_i$ from agent i. Here $f_i > 0$ is polluter i's tax rate. The Pigouvian fee is $f_i = \sum_{j \in R^0 i} a_{ij}$ for every polluter $i \in N$. It implements the first-best emissions e^* in Nash equilibrium. Alternatively, the fee can be on ambient pollution rather

than on emissions. A pollution fee scheme t^F charges $F_j > 0$ per unit of emissions at each receptor j which leads agent i to pay $t_i^F(e) = -\sum_{j \in Ri} a_{ij} F_j e_i$. The emission or ambient pollution fee mechanism can be associated with a redistribution policy of the money collected, e.g. through lump-sum transfers or subsidies.

A further important regulation instrument that can be embedded in our model is cap-and-trade or tradable emission permits. Agents are endowed with some initial emissions allowances or permits $\bar{e} = (\bar{e}_i)_{i \in N}$ which can be traded in a market. They are not allowed to emit more than the amount of permits they own at the end of a pre-pollution trading phase. Providing that the permit market is competitive (implying that agents are price takers), the tradable emission permit regulation is as if each agent i faces a transfer scheme $t_i^{tp}(e) = p(\bar{e}_i - e_i)$ where p is the equilibrium price of permits. This price is uniquely determined by the first-order conditions $b_j'(e_j^t) - a_{jj} = p$ for every $j \in N \backslash V$ and the market clearing condition $\sum_{j \in N} \bar{e}_j = \sum_{j \in N} e_j^t$. The initial allocation of permits impacts the level and distribution of welfare. Under grandfathering, each agent is assigned a share of his or her laissez-faire emission $\bar{e}_i = \alpha e_i^{lf}$ with $0 < \alpha \leq 1$. A lower α means lower emissions in the economy. When permits are auctioned by the government, it is as if those who get the revenue from this auction are endowed with the permits. For instance, if the money is used exclusively to reduce or compensate the damage at agent h's location, then it is as if agent h obtains all permits and trades them with polluters in a competitive market, i.e. $\bar{e}_h = \sum_{j \in N \backslash V} e_j^t$. Emission allowances can also be defined on receptors emissions, each agent i potentially owning $\bar{e}_{ij}$ emission allowances at receptor j that can be exchange against other emission allowances for the same receptor j.

Given the abundance of different regulation mechanisms in reality, a society would like to distinguish between them according to desirable criteria. The following will be two very basic requirements any society would like any regulation to comply with.

Efficiency requires that the first-best outcome is implemented in Nash equilibrium.

Efficiency: For all problems a and all $e^t \in \mathcal{N}(t, a)$, we have $e^t = e^*$.

The second property requires that the payments of the mechanism are budget-balancing at Nash equilibrium.

Budget Balance: For all problems a and all $e^t \in \mathcal{N}(t,a)$, we have $\sum_{i\in N} t_i(e^t) \leq 0$.

A budget balanced regulation mechanism t where $\mathcal{N}(t,a)$ is a singleton for any a, say $\mathcal{N}(t,a) = \{e^t\}$, induces a distribution rule ϕ^t of the total welfare. For any problem a, the distribution rule implemented by the budget balanced mechanism t sets:

$$\phi_i^t(a) = b_i(e_i^t) - \sum_{j\in Si} a_{ji} e_j^t + t_i(e^t).$$

Any of the above regulation mechanisms is budget balanced and has a unique Nash equilibrium, and hence, induces a corresponding distribution rule. We now focus on a particular regulation mechanism, the one inspired by the polluter-pays principle.

4 A Characterization of the Polluter-Pays principle

In this section, we first describe the polluter-pays mechanism and show two of its properties, namely budget balancedness and efficiency. Second, we examine the properties of the welfare distribution rules implemented by regulation mechanisms in Nash equilibrium, and, in particular by the polluter-pays welfare distribution rule.

4.1 The Polluter-Pays Mechanism

Many countries have adopted the "polluter-pays" (PP) principle as a regulation mechanism. It basically renders the polluter responsible for the damage it causes to the environment. It requires that *the costs of pollution should be borne by the entity which profits from the process that causes pollution.* In order to satisfy the polluter-pays principle, the entity who pollutes should compensate those who suffer from this pollution for the damages it causes. If a victim is not fully compensated then he or she pays part of the cost of someone else's pollution. Hence, strictly speaking, the PP principle imposes not only that polluters pay for the damage caused to society, but also that victims are fully compensated for those damages. In our model, an

arbitrary agent i who pollutes should compensate every agent $j \in R^0 i$ for the caused damage $a_{ij}e_i$. Agent i pays $a_{ij}e_i$ to every $j \in R^0 i$. Therefore, as a victim of pollution, agent i receives the compensation $a_{ji}e_j$ from each agent $j \in S^0 i$ who pollutes him. Summing up all these side-payments, the polluter-pays principle leads to the regulation mechanism $t^{PP}(e)$ defined as follows for any agent $i \in N$:

$$t_i^{PP}(e) = \sum_{j \in S^0 i} a_{ji}e_j - \sum_{j \in R^0 i} a_{ij}e_i = d_i - a_{ii}e_i - \sum_{j \in R^0 i} a_{ij}e_i = d_i - \sum_{j \in Ri} a_{ij}e_i. \tag{4}$$

Agent i receives the net transfer from the cost of pollution he suffers minus the cost of pollution he causes to society. Since the polluter-pays principle involves side-payments among agents, the payments in the PP-mechanism sum up to zero. It is therefore budget-balanced. Agent i's welfare under the payments $t^{PP}(e)$ with emission plan e is:

$$b_i(e_i) - \sum_{j \in Ri} a_{ij}e_i \tag{5}$$

Since agent i pays for the marginal damage caused to others and is compensated from the marginal damage caused by others, his welfare under the PP-mechanism in (5) is the social benefit from his economic activity. Therefore, agent i has incentive to emit the efficient level e_i^* for any given emissions emitted by other agents. Formally, maximizing (5) with respect to e_i leads to the first-order condition (2) which implies $e_i^t = e_i^*$ for every $i \in N$. This implies that the PP-mechanism implements the efficient emission plan e^* in Nash equilibrium, i.e. $\mathcal{N}(t^{PP}, a) = \{e^*\}$. A particular feature of regulation through the PP-mechanism with constant marginal damages is that, since any individual's payoffs depend only on the agent's own choice (no externality), the efficient emission plan is a dominant strategy equilibrium, which is an equilibrium concept which is less demanding in terms of cognitive skills than Nash equilibrium. Therefore, the efficient emissions plan remains the unique Nash equilibrium when the parameters a are publicly known but the benefit functions are private information. One can even check that the efficient emissions plan is robust to collusion, i.e. it remains the unique equilibrium in the PP mechanism even if we allow coalitions to jointly change their emissions.[7]

[7]This is easily seen by the following argument: for any non-empty coalition $S \subseteq N$ we have that $(e_i^*)_{i \in S}$ solves $\max_{(e_i)_{i \in S} \geq 0} \sum_{i \in S} [b_i(e_i) + t_i^{PP}((e_j)_{j \in S}, (e_j^*)_{j \in N \setminus S})]$.

We will denote by ϕ^{PP} the polluter-pays (PP) distribution rule associating with each problem (N, b, a). Its polluter-pays welfare distribution $\phi^{PP}(a)$ is given by agent i's equilibrium welfare for every $i \in N$:

$$\phi_i^{PP}(a) = b_i(e_i^*) - \sum_{j \in Ri} a_{ij} e_i^*. \tag{6}$$

The result below follows straightforwardly from our discussion.

Proposition 1 *The polluter-pays mechanism is an efficient and budget-balanced regulation implementing the polluter-pays distribution rule.*

4.2 The Polluter-Pays Distribution Rule

The following are two desirable criteria a society would like to be satisfied by the welfare distributions implemented via a regulation mechanism. The first criterium requires that any agent should receive a non-negative payoff.

Non-Negativity: For all problems a and all $i \in N$, $\phi_i(a) \geq 0$.

In the absence of pollution or emission activities, any agent's welfare is zero and the state of no pollution may be interpreted as status quo. Non-negativity simply requires that nobody should be worse off under pollution than without pollution.

The second criterium renders the polluter responsible to any change of his pollution impact on the economy.

Responsibility for Pollution Impact (RPI): Consider any arbitrary agent $i \in N$. Suppose that agent i's pollution impact is reduced from $(a_{ij})_{j \in N}$ to $(a'_{ij})_{j \in N}$ with $a_{ij} \geq a'_{ij}$ for all $j \in N$, and all other pollution impacts being unchanged ($a'_{lj} = a_{lj}$ for all $l \in N \backslash \{i\}$ and all $j \in N$). The distribution rule ϕ renders agents responsible for their pollution impact if for any $i \in N$, any reduction a' of i's pollution impact from a,

$$\phi_i(a') - \phi_i(a) = W(a') - W(a).$$

Responsibility for pollution impact (RPI) requires to assign to any agent the full return or loss of any change of his own pollution impact.

In addition to being a fairness principle, RPI has attractive incentive properties. Suppose that an agent is able to reduce his pollution impact at some cost by switching to a greener technology, reducing or cleaning its wastes, improving energy efficiency or using less toxic inputs. By assigning the full return of this pollution reduction, RPI provides efficient incentives to invest in pollution impact reduction. Symmetrically, if an agent benefits from increasing his pollution impact per unit of emissions (e.g. using higher sulfur content coal), RPI assigns to this agent the economic cost of this extra pollution.

Among the above regulations, the Pigouvian fee regulation mechanism is efficient. It is budget balanced if the revenue collected is redistributed to agents. The welfare distribution it implements does not satisfy non-negativity since victim-only agents (i.e. agents $i \in V$) are not compensated for the environmental damage they incur. The welfare distribution with the Pigouvian fee regulation mechanism also satisfies RPI. An emission norm $\bar{e}_i = e_i^*$ with a persuasive fine (e.g. infinite) is efficient and budget balanced but its welfare distribution does not satisfy RPI and non-negativity. A cap-and-trade system (tradable pollution allowances) for pollution at each receptor with grandfathering is efficient and budget balanced but the welfare distribution it leads to does not satisfy non-negativity since victims are not compensated entirely. It might or might not satisfy RPI depending on the initial allocation of permits. A similar cap-and-trade system where permits are auctioned satisfies efficiency and RPI but it is not budget balanced unless the money collected is redistributed.

Theorem 1 *The polluter-pays distribution rule is the unique distribution rule that satisfies non-negativity and responsibility for pollution impact.*

Proof. First, we show that if a distribution rule satisfies non-negativity and responsibility for pollution impact, then it must be the polluter-pays distribution rule ϕ^{PP}. Consider another distribution rule ϕ and let a be a problem. Let $\phi(a) = \tilde{z}$ and $\phi^{PP}(a) = z^{PP}$. Let $\sum_{i\in N} \tilde{z}_i = \tilde{W}$. Suppose that $\tilde{z} \neq z^{PP}$. Since $\sum_{i\in N} z_i^{PP} = W(a)$ and $\tilde{z}$ is a welfare distribution, we have $\tilde{W} \leq W(a)$. Thus, $\sum_{i\in N} \tilde{z}_i \leq \sum_{i\in N} z_i^{PP}$ which, combined $\tilde{z} \neq z^{PP}$ forces $\tilde{z}_i < z_i^{PP}$ for some $i \in N$. Note that for all $j \in V$, $z_j^{PP} = 0$ and by non-negativity of ϕ, $\tilde{z}_j \geq 0$. Thus, we must

have $i \in N \backslash V$ and both $a_{ii} > 0$ and $\hat{e}_i > 0$. Let a' be such that a is a pollution impact reduction for agent i from a' such that $a_{ii} < a'_{ii}$ and everything else remains identical, i.e. $a'_{lj} = a_{lj}$ for all $l, j \in N$ such that $lj \neq ii$. Pick a'_{ii} sufficiently large such that

$$b_i(e_i'^{lf}) < z_i^{PP} - \tilde{z}_i \tag{7}$$

where $\mathcal{N}(t^{lf}, a') = \{e'^{lf}\}$. Let $\phi(a') = \tilde{z}'$ and $\phi^{PP}(a') = z'^{PP}$ denote the distributions chosen by ϕ and ϕ^{PP} for the problem (N, b, a'). By responsibility for pollution impact,

$$\tilde{z}_i - \tilde{z}'_i = z_i^{PP} - z_i'^{PP}$$

Rearranging terms and using the definition of z_i^{PP} this leads to

$$z_i^{PP} - \tilde{z}_i = b_i(e_i'^*) - \sum_{j \in Ri} a'_{ij} e_i'^* - \tilde{z}'_i, \tag{8}$$

where e'^* denotes the efficient emission plan for (N, b, a'). By non-negativity of ϕ, $\tilde{z}'_i \geq 0$. Now since $b_i(e_i'^{lf}) \geq b_i(e_i'^*)$, $a'_{ij} \geq 0$ for all $j \in Ri$, and $z'_i \geq 0$, we obtain from (7),

$$z_i^{PP} - \tilde{z}_i > b_i(e_i'^{lf}) \geq b_i(e_i'^*) - \sum_{j \in Ri} a'_{ij} e_i'^* - \tilde{z}'_i,$$

which contradicts (8).

Second, we show that ϕ^{PP} satisfies non-negativity and responsibility for pollution impact. For non-negativity,

$$z_i^{PP} = b_i(e_i^*) - \sum_{j \in Ri} a_{ij} e_i^* = \max_{e_i \geq 0} \left(b_i(e_i) - \sum_{j \in Ri} a_{ij} e_i \right) \geq b_i(0) - \sum_{j \in Ri} a_{ij} \times 0 = 0,$$

where the inequality follows from the fact that agent i can always choose $e_i = 0$ (no emission or production).

For responsibility for pollution impact, for any agent i, consider any reduction of i's pollution impact from a to a': $a_{ij} \geq a'_{ij}$ for all $j \in N$ and $(a_{kj})_{j \in N} = (a'_{kj})_{j \in N}$ for any $k \neq i$. Let $\phi^{PP}(a) = z^{PP}$ and $\phi^{PP}(a') = z'^{PP}$. Let $W(a)$ and $W(a')$ denote the corresponding total welfare in (N, b, a) and (N, b, a'), respectively. Note that by efficiency of t^{PP}, we have both $W^{PP}(a) = W(a)$ and $W^{PP}(a') = W(a')$. Similarly, denote by e^* and e'^* the efficient emission plan of (N, b, a) and (N, b, a'), respectively. By definition,

$$z_i'^{PP} - z_i^{PP} = b(e_i'^*) - \sum_{j \in Ri} a'_{ij} e'^*_{ij} - \left(b(e_i^*) - \sum_{j \in Ri} a_{ij} e^*_{ij} \right). \tag{9}$$

Since $a_{kj} = a'_{kj}$ for every $k \neq i$, the efficient emission levels are not affected by the change of matrix of pollution impacts from a to a' which implies $e^*_k = e'^*_k$ for every $k \in N \backslash i$. Therefore, we have:

$$W(a') - W(a) = b(e'^*_i) - \sum_{j \in Ri} a'_{ij} e'^*_{ij} - \left(b(e^*_i) - \sum_{j \in Ri} a_{ij} e^*_{ij} \right)$$

which, combined with (9), leads to $z'^{PP}_i - z^{PP}_i = W'^{PP} - W^{PP}$. □

Because for any problem a, $\phi^{PP}(a)$ is an efficient welfare distribution, Theorem 1 shows that non-negativity and responsibility for pollution impact imply efficiency.

5 Acceptability of the polluter-pays principle

In modern societies, environmental regulations emerge from negotiation among stakeholders. Sovereign countries bargain to design environmental international agreements such as the Kyoto protocol. Firms, public authorities and NGO are involved in the debates on the design of regulations. In this section, we analyze such negotiations using a cooperative game theory approach. We examine whether the polluter-pays mechanism is the preferred mechanism for any possible coalition of agents. That is if a group of agents can be better-off by agreeing on another way to regulate externalities among them and leaving out the other agents. For instance, in the case of international agreements, a group of countries would refuse to agree to apply the polluter-pays mechanism if it can achieve a higher welfare with another agreement among them. Alternatively, some countries may refuse to sign a global agreement and only a subcoalition of the grand coalition signs the agreement and cooperates while the non-signatory countries choose their emissions non-cooperatively.

Our analysis requires additional notation from cooperative game theory. Throughout this section we suppose that (N, b, a) is fixed and $z^{PP} = \phi^{PP}(a)$. A coalition T is a non-empty subset of N. For any coalition T, let $e_T = (e_i)_{i \in T}$. We need to define the welfare that a coalition T can achieve by agreeing on a regulation mechanism. This welfare depends on the behavior of agents outside the coalition. We assume that if a coalition T forms, then the members of T agree to implement the mechanism that maximizes their joint payoff given the

behavior of agents outside T. As outlined above, here we assume that agents outside of T behave non-cooperatively.[8] For our purpose of computing T's welfare, this is equivalent to agree on an emission plan e_T among members of T. Agents outside T simply choose their laissez-faire emissions.

More precisely, for any coalition T and emissions e'_j for agents $j \in N\backslash T$ outside T, members of coalition T would implement the solution to:

$$\max_{(e_i)_{i\in T}\in\mathbb{R}_+^T} \sum_{i\in T}\left(b_i(e_i) - \left(\sum_{j\in T} a_{ji}e_j + \sum_{j\in N\backslash T} a_{ji}e'_j\right)\right).$$

The first-order conditions defines the emissions e^T of members of coalition T:

$$b'_i(e_i^T) = \sum_{j\in Ri\cap T} a_{ij}, \tag{10}$$

for all $i \in T$, and we set $e_i^T = e_i^{lf}$ for all $i \in N\backslash T$. Importantly, with constant marginal damage, coalition T's emissions do not depend on the outsider emissions $(e'_j)_{j\in N\backslash T}$. In particular, the members of T choose the same emissions if the others agents coordinate their emissions (by forming coalitions) or not. Each agent $i \in T$ internalizes his impact on the environmental damage only to members of T. Therefore, for any $i \in T$, e_i^T weakly decreases as T expands.

For each coalition T, e^T defined by (10) determines the welfare T can achieve:

$$\underline{v}(T) = \sum_{i\in T}\left(b_i(e_i^T) - \left(\sum_{j\in T} a_{ji}e_j^T + \sum_{j\in N\backslash T} a_{ji}e_j^T\right)\right).$$

We call $\underline{v}(T)$ the non-cooperative core lower bound for coalition T. This is the maximum welfare T can achieve by cooperation while agents outside T behave non-cooperatively.

Non-Cooperative Core Lower Bounds: A welfare distribution z satisfies non-cooperative core lower bounds if for every $T \subseteq N$, $\sum_{i\in T} z_i \geq \underline{v}(T)$.

Not surprisingly, the polluter-pays welfare distribution does not necessarily satisfy non-cooperative lower bounds because big polluters might decide to stay alone even when all other agents behave non-cooperatively.

[8] That is to say they do not form any cooperating coalitions of size greater than or equal to two.

Proposition 2 *The polluter-pays welfare distribution z^{PP} might not satisfy non-cooperative core lower bounds.*

Proof. First, note that for any agent i,

$$z_i^{PP} = b_i(e_i^*) - \sum_{j \in Ri} a_{ij} e_i^* = b_i(e_i^*) - a_{ii} e_i^* - \sum_{j \in R^0 i} a_{ij} e_i^*$$

and the non-cooperative core lower bound for $T = \{i\}$ is

$$\underline{v}(i) = b_i(e_i^{lf}) - a_{ii} e_i^{lf} - \sum_{j \in S^0 i} a_{ji} e_j^{lf}$$

Since e_i^{lf} maximizes $b_i(e_i) - a_{ii} e_i$, as long as agent i exerts some externalities, i.e. $R^0 i \neq \emptyset$, a sufficient condition for the core lower bound for coalition $\{i\}$ to be violated, i.e. for $z_i^{PP} < \underline{v}(i)$, is

$$\sum_{j \in R^0 i} a_{ij} e_i^* \geq \sum_{j \in S^0 i} a_{ji} e_j^{lf}. \tag{11}$$

Obviously, (11) holds if agent i is not polluted by others, i.e. $S^0 i = \emptyset$. More generally, it holds when the damage from others at the laissez-faire is lower than the damage to others at the first-best. In this case, the payment imposed to agent i by the polluter-pays principle exceeds the laissez-faire cost of pollution for i. One can easily find examples where (11) is violated with unilateral externalities. With multilateral externalities $S^0 i = R^0 i$ and, therefore, (11) becomes:

$$\sum_{j \in S^0 i} a_{ij} e_i^* \geq \sum_{j \in S^0 i} a_{ji} e_j^{lf}.$$

The above condition holds if i has a large marginal impact on the others while the others have a low impact on i. Then i is required to pay a lot when his gain is low. Here also we can find examples where the last inequality is violated with multilateral externalities. □

Proposition 2 and its proof implies that some countries might not sign a global agreement because they prefer to stay alone (while some other coalition forms). For example, a country might refuse to ratify an international agreement for emissions reduction (e.g. the Kyoto Protocol) because its impact on others is much larger than the pollution it suffers from others.

This is exactly captured in (11). Also note that for the above conclusions for such a country it is irrelevant whether all countries behave non-cooperatively or a coalition forms and all others behave non-cooperatively: if some coalition cooperates, then their emissions are reduced from laissez-faire and the pollution impact is smaller on this country (like the United States) than under laissez-faire. More precisely, a country prefers all other countries to sign the international environmental agreement while they do not sign it!

Interestingly, the polluter-pays welfare distribution satisfies all non-cooperative core lower bounds in a symmetric pollution environment, i.e. where all benefit functions are identical and pollution impacts are the same across all agents.

Proposition 3 *Under multilateral externalities and homogenous agents where for all $i, j \in N$, $b = b_i = b_j$ and $a = a_{ij} = a_{ji}$, the polluter-pays welfare distribution z^{PP} satisfies the non-cooperative core lower bounds.*

Proof. For all $i, j \in N$, let $b = b_i = b_j$ and $a = a_{ij} = a_{ji}$. Since $a_{ii} > 0$ for all $i \in N$, we have $Si = Ri = N$ for all $i \in N$. Let $\emptyset \neq T \subseteq N$.

Now for all $i \in T$,

$$\begin{aligned}
z_i^{PP} &= b(e_i^*) - \sum_{j \in N} a e_i^* \\
&= b(e_i^*) - e_i^* \sum_{j \in N} a \\
&= \max_{e_i \geq 0} \left(b(e_i) - e_i \sum_{j \in N} a \right) \\
&\geq b(e_i^T) - e_i^T \sum_{j \in N} a \\
&= b(e_i^T) - \sum_{j \in T} a e_i^T - \sum_{j \in N \setminus T} a e_i^T \\
&\geq b(e_i^T) - \sum_{j \in T} a e_j^T - \sum_{j \in N \setminus T} a e_j^T .
\end{aligned}$$

where the first equality is the definition of the polluter-pays welfare distribution, the second follows from the fact that in the homogenous case, for all $j \in N$, $e_i^* = e_j^*$, and the last inequality follows from the fact that in the homogenous case, for all $j \in T$, $e_i^T = e_j^T$, and that for all $j \in N \setminus T$, $e_i^T \leq e_i^{lf} = e_j^{lf} = e_j^T$.

Now the above implies

$$\sum_{i\in T} z_i^{PP} \geq \sum_{i\in T}\left(b(e_i^T) - \sum_{j\in T} a e_j^T - \sum_{j\in N\setminus T} a e_j^T\right) = \underline{v}(T).$$

Hence, z^{PP} satisfies non-cooperative core lower bounds, the desired conclusion. □

Proposition 3 shows that in homogenous environments no coalition has an incentive to leave a global agreement where the polluter-pays principle is applied. As in real life applications, society may decide that a global agreement is only enforced if all agents sign it (and otherwise laissez-faire or a partial agreement is adopted). In the homogenous case all agents weakly prefer signing the global agreement instead of discarding it.

6 Generalization to increasing marginal damage

We now consider the polluter-pays principle with convex damage functions which requires a slight modification of the model. We differentiate emissions from pollution and damage. The emission plan e generates a pollution level p_i at i's location (to receptor i) defined by:

$$p_i = \sum_{j\in Si} a_{ji} e_j. \tag{12}$$

The matrix a defines now the *transfer coefficients* that translates emissions of i into pollution of j (e.g. waste water released by i into water pollution concentration on j). Pollution at level p_i causes damages $d_i(p_i)$ to i with d_i being increasing and convex: $d_i(0) = 0$, $d_i'(p_i) > 0$ and $d_i''(p_i) \geq 0$ for every $p_i \in \mathbb{R}_+$ and $i \in N\setminus P$.[9] The welfare of agent i with emissions $e = (e_i)_{i\in N}$ is:

$$b_i(e_i) - d_i(p_i), \tag{13}$$

where p_i is defined by (12). A pollution problem is now described by (N, b, a, d).

[9] Recall that P is the set of only polluter agents.

The first-order conditions that characterize the efficient emission plan e^* (which maximizes the total welfare $\sum_{i\in N}[b_i(e_i) - d_i(p_i)]$) are for every $i \in N$:[10]

$$b_i'(e_i^*) = \sum_{j\in Ri} a_{ij} d_j'(p_j^*) = \sum_{j\in Ri} a_{ij} d_j'(\sum_{l\in Sj} a_{lj} e_l^*). \tag{14}$$

The marginal benefit of agent i's emission should be equal to its marginal cost for society which depends on its marginal impact on pollution a_{ij} and the marginal damage of pollution at each receptor $j \in Ri$. Each unit of emission from agent i leads to a_{ij} units of pollution at receptor j which causes marginal damages evaluated to $a_{ij}d_j'(p_j^*)$. The total welfare with the efficient emission plan e^* is:

$$W(a) = \sum_{i\in N}[b_i(e_i^*) - d_i(p_i^*)] = \sum_{i\in N}[b_i(e_i^*) - d_i(\sum_{j\in Si} a_{ji} e_j^*)].$$

In contrast with constant marginal damages (i.e. the first-order condition in (2)), with increasing marginal damage the efficient level of i's emission (the first-order condition in (14)) depends on what is emitted by the other polluters of j with j being a receptor of i's pollution ($j \in Ri$). Marginal damage being increasing with pollution concentration, agent i's emission has more impact on damages at j when pollutant emitted by other polluters in $R^0 j$ increases. Because a polluter's marginal impact depends on pollution concentration due to other polluters, applying the polluter-pays principle in this framework is not straightforward. One needs to define each polluter's responsibility on the damage caused to society when computing the "cost of pollution of one entity on others". With only one single polluter i, it is easy: agent i should pay the damage $d_j(a_{ij}e_j)$ to victim j. However, with more than one polluter at a receptor j, say i and k, the PP principle does not tell us how to share $d_j(a_{ij}e_i + a_{kj}e_k)$ (the overall cost at j) among i and k. If polluter i is held responsible for the first $a_{ij}e_j$ units of pollution, he has to pay $d_j(a_{ij}e_i)$. If polluter i is responsible for the last ones, he has to pay $d(a_{ij}e_i + a_{kj}e_k) - d_j(a_{kj}e_k)$ which is larger than $d_j(a_{ij}e_i)$ by convexity of d_j. It is also

[10] The existence of the efficient emission plan e^* is guaranteed by Brouwer's fixed point theorem: define $g : \times_{i\in N}[0, \hat{e}_i] \longrightarrow \times_{i\in N}[0, \hat{e}_i]$ by $g(e) = ((b_i')^{-1}(\sum_{j\in Ri} a_{ij} d_j'(\sum_{l\in Sj} a_{lj} e_l)))_{i\in N}$. Since b_i is strictly concave, b_i' tends to infinity at zero, and b_i' tends to zero at $\hat{e}_i$, g is a well defined function. Our assumptions on damages ensure that g is continuous. Now since $\times_{i\in N}[0, \hat{e}_i]$ is compact and convex, Brouwer's fixed point theorem implies that the function g must have a fixed point which is a solution to (14). Uniqueness of e^* follows from strict concavity of the benefits and the convexity of the damages.

increasing with the other polluter k's emissions. One can think about several ways to share the damage $d_j(p_j)$. For instance, it could be assigned proportionally to a polluter's share on total pollution, each polluter i paying $\frac{a_{ij}e_i}{p_j}d_j(p_j)$ to j for every $i \in Rj$.

Such a division of the damage is defined for given emissions by i and k. Yet, since emissions are substitutes for receptor j, the presence of i's emissions at j leads to a reduction of k's emissions e_k at the first-best. The inter-connection of polluters' efficient emissions with convex damage creates a further cost of pollution on society: i's emission do not only cause damage at j, it also encroaches on k's emission at the first-best.

In this framework, we interpret the PP principle of making paying the "cost of pollution of one entity on others" by charging a polluter the incremental impact of his emissions on other agents. Due to increasing marginal damage, we can distinguish between two impacts. A first one is an increase of damage at each receptor $j \in Ri$. The second one is due to the substitution between polluters' emissions for each receptor j: if i emits more pollution, then each polluter $k \in Sj$ should emit less at the first-best. We also interpret the PP principle by compensating each agent exactly for the damage caused by others' emissions in absence of his emission. Let us denote by e^{0i} the efficient emission *without* i's emission for every $i \in N$ (with fixing $e_i^{0i} = 0$). Notice that e^{0i} is the efficient plan of an economy *without* i's emission but *with* i's damage function d_i (i.e. agent i is then a "victim only"). It maximizes the total welfare of the problem (N, b_{-i}, a, d) where by b_{-i} implicitly means that agent i becomes a victim. The polluter-pays regulation mechanism t^{PP} is defined for every $i \in N$ by:

$$t_i^{PP}(e) = d_i(p_i^{0i}) - \sum_{j \neq i} \left[b_j(e_j^{0i}) - d_j(p_j^{0i}) - (b_j(e_j) - d_j(p_j)) \right]. \tag{15}$$

The transfer is decomposed in two terms. The left-hand term is agent i's damage at the first-best without i's emissions. The summation is the economic loss due to i's emission for all other agents. A polluter j who is victim of i's pollution, the transfer is simply the loss of benefit $b_j(e_j^{0i}) - b_j(e_j)$. For a polluter j who is a victim of i's pollution it is the change of welfare including damage $b_j(e_j^{0i}) - d_j(p^{0i}) - (b_j(e_j) - d_j(p_j))$. For a victim only agent $i \in V$, the PP transfer reduces to the first term which is the damage at the first-best $d_i(p_i^*)$.

The PP mechanism yields to agent i a total welfare of (noting $b_i(e_i^{0i}) = b_i(0) = 0$):

$$b_i(e_i) - d_i(p_i) + t_i^{PP}(e) = \sum_{j \in N} \left[b_j(e_j) - d_j(p_j) - (b_j(e_j^{0i}) - d_j(p_j^{0i})) \right]. \tag{16}$$

Agent i's welfare under the PP regulation mechanism is his emission's contribution to total welfare for any emission plan. Since each agent i internalizes the impact of his own emissions on total welfare given the other agent's emissions, the PP principle implements the efficient emission plan e^*. Indeed, given other agent's emissions e^t_{-i}, the maximization of agent i's welfare

$$b_i(e_i) - \sum_{j\in Ri} d_j(a_{ij}e_i + \sum_{l\in Sj\setminus i} a_{lj}e^t_j) + \sum_{j\in N\setminus i} b_j(e^t_j) - \sum_{j\in N\setminus Ri} d_j(p^t_j) - \sum_{i\in N}(b_j(e^{0i}_j) - d_j(p^{0i}_j))$$

with respect to e_i leads to the first-order conditions in (14) of the efficient emission plan e^*. Therefore, $e^t = e^*$ for any $e^t \in \mathcal{N}(t,a)$. Agent i's equilibrium welfare is thus by (16):

$$\phi^{PP}_i(a) = b_i(e^*_i) - d_i(p^*_i) + t^{PP}_i(e^*) = W(a) - \sum_{j\in N}(b_j(e^{0i}_j) - d_j(p^{0i}_j)). \tag{17}$$

where $W(a) = \sum_{j\in N}(b_j(e^*_j) - d_j(p^*_j)) = W^{PP}(a)$. Similarly as before, we call ϕ^{PP} the polluter-pays distribution rule (induced by t^{PP} for convex damages). Agent i's welfare is the incremental contribution of his emissions at the first-best. For a victim only agent $i \in V$, it simplifies to zero since he is fully compensated for the damage $d_i(p^*_i)$. A polluter only agent $i \in P$ obtains his first-best benefit $b_i(e^*_i)$ net of the negative impact of his emissions on society $\sum_{j\neq i}\left[b_j(e^*_j) - d_j(p^*_j) - (b_j(e^{0i}_j) - d_j(p^{0i}_j))\right]$. Note that since $e^{0i}_j = e^*_j$ with constant marginal damages for every $j \neq i$ the PP mechanism defined in (15) is a generalization of the one defined in (4) to convex damage functions. The next proposition shows that t^{PP} is budget-balanced.

Proposition 4 *The polluter-pays mechanism is an efficient and budget-balanced regulation implementing the polluter-pays distribution rule.*

Proof. It remains to be shown that t^{PP} is budget-balanced. Since t^{PP} is efficient, it suffices to show $\sum_{i\in N} t^{PP}_i(e^*) \leq 0$. Note that since e^{i0} is an efficient emission plan of the problem (N, b_{-i}, a, d) while the emission plan $(e^*_{-i}, 0)$ can be implemented in (N, b_{-i}, a, d), we have

$$\sum_{j\in N}\left[b_j(e^{0i}_j) - d_j(p^{0i}_j)\right] \geq -d_i(p^*_i - a_{ii}e^*_i) + \sum_{j\neq i}\left[b_j(e^*_j) - d_j(p^*_j - a_{ij}e^*_i)\right].$$

Multiplying both sides with -1, we combine the above inequality with the definition of $t^{PP}(e)$ in (15) at the first-best and use $b_i(e^{0i}_i) = b_i(0) = 0$ and $a_{ij} = 0$ for $j \notin Ri$, and obtain:

$$t^{PP}_i(e^*) \leq d_i(p^*_i - a_{ii}e^*_i) - \sum_{j\in R^0 i}\left[d_j(p^*_j) - d_j(p^*_j - a_{ij}e^*_i)\right]$$

Summing up all transfers t_i^{PP} leads to:

$$\sum_{i\in N} t_i^{PP}(e^*) \leq \sum_{i\in N}(d_i(p_i^* - a_{ii}e_i^*) - \sum_{j\in R^0 i}\left[d_j(p_j^*) - d_j(p_j^* - a_{ij}e_i^*)\right])$$

Rearranging terms yields:

$$\sum_{i\in N} t_i^{PP}(e^*) \leq \sum_{i\in N}(d_i(p_i^* - a_{ii}e_i^*) - \sum_{j\in S^0 i}[d_i(p_i^*) - d_i(p_i^* - a_{ji}e_j^*)]). \tag{18}$$

Consider any $i \in N$. Without loss of generality, let $Si = \{1, \ldots, s\}$. Since $p_i^* = \sum_{j\in Si} a_{ji}e_j^*$, we can rewrite $d_i(p_i^*)$ by:

$$d_i(p_i^*) = \sum_{k=1}^{s}[d_i(p_i^* - \sum_{j=1}^{k-1} a_{ji}e_j^*) - d_i(p_i^* - \sum_{j=1}^{k} a_{ji}e_j^*)] \tag{19}$$

Note that for any $k = 1, \ldots, s$, $p_i^* - \sum_{j=1}^{k-1} a_{ji}e_j^* - (p_i^* - \sum_{j=1}^{k} a_{ji}e_j^*) = a_{ki}e_k^* = p_i^* - (p_i^* - a_{ki}e_k^*)$ Thus, by convexity of d_i, for any $k = 1, \ldots, s$,

$$d_i(p_i^* - \sum_{j=1}^{k-1} a_{ji}e_j^*) - d_i(p_i^* - \sum_{j=1}^{k} a_{ji}e_j^*) \leq d_i(p_i^*) - d_i(p_i^* - a_{ki}e_k^*) \tag{20}$$

Combining (19) with (20) for any $k = 1, \ldots, s$ leads to:

$$d_i(p_i^*) \leq \sum_{k=1}^{s}[d_i(p_i^*) - d_i(p_i^* - a_{ki}e_k^*)] = \sum_{j\in Si}[d_i(p_i^*) - d_i(p_i^* - a_{ji}e_j^*)].$$

By $Si = S^0 i \cup \{i\}$, this is equivalent to:

$$d_i(p_i^* - a_{ii}e_i^*) \leq \sum_{j\in S^0 i}\left[d_i(p_i^*) - d_i(p_i^* - a_{ji}e_j^*)\right].$$

The last inequality combined with (18) leads to the desired conclusion. □

Notice that as along as two polluters impact the same receptors, the PP distribution rule does not distribute total welfare in the sense that $\sum_{i\in N} \phi_i^{PP}(a) < W(a)$. To see that, suppose that $N = \{1, 2, 3\}$ with polluter 1 and 3 polluting only a victim 2, i.e. $a_{i2} > 0$ for $i = 1, 3$. Then polluter 1 has to pay the incremental damage at 2, formally $d_2(a_{12}e_1^* + a_{32}e_3^*) - d_2(a_{32}e_3^{01})$ as well as the loss of benefit for 3, that is $b_3(e_3^{01}) - b_3(e_3^*)$. Similarly polluter 3 has to pay for increment damages at 2 and benefit loss for 1 due to his emissions. The victim 2 receives a compensation equals to the damage $d_2(p_2^*)$. Yet, the total payment by 1 and 3 more than

offsets the compensation to 2: $t_1^{PP}(e^*)+t_3^{PP}(e^*)+t_2^{PP}(e_2^*)<0$ because $-t_1^{PP}(e^*)-t_3^{PP}(e^*)>t_2^{PP}(e_2^*)=d_2(a_{12}e_1^*+a_{32}e_3^*)$. The PP regulation mechanism exhibits a financial surplus and, therefore, the PP regulation rule distributes strictly less than total welfare.

To characterize the PP distribution rule ϕ^{PP} with marginal increasing damages, we introduce a further fairness principle, called solidarity upper bounds. Its motivation relies on polluters' minimal claims when applying the PP principle. To minimize his payment, a polluter would claim responsibility only on the damage impact due to his own emission in absence of any other pollution at each receptor $j \in Ri$ (including himself). Under this interpretation of the PP principle each agent i would enjoy an individual welfare of $\max_{e_i\geq 0}[b_i(e_i)-\sum_{j\in Ri} d_j(a_{ij}e_j)]$ for a given emission plan e. On the other hand, applying the PP principle requires to fully compensate any agent j for the damage $d_j(p_j)$. With (strictly) convex damage functions we have $\sum_{i\in Sj} d_j(a_{ij}e_i) < d_j(p_j) = d_j\left(\sum_{i\in Si} a_{ij}e_i\right)$ whenever $|Si|>1$, and such an interpretation of the polluter pays principle would lead to unbalanced transfers. One way to reconcile a distribution rule (or budget-balanced transfers) with the above claims is to impose that, by solidarity, no agent should get more than the claimed stand-alone welfare. This solidarity principle is given in the following requirement.

Solidarity Upper Bounds: For all problems a and all $i \in N$, $\phi_i(a) \leq \max_{e_i\geq 0}[b_i(e_i) - \sum_{j\in Ri} d_j(a_{ij}e_i)]$.

A second, and more fundamental, justification of solidarity upper bounds finds its roots in Moulin's notion of group externality (Moulin, 1990). Under increasing marginal damage, the presence of pollution from other sources might reduce the ability of a polluter to emit. Formally, let us denote by e_i^{0-i} polluter i's efficient emission when i is the only polluter to emit ($e_j = 0$ for every $j \neq i$). It is the efficient emission plan to the problem (N, b_i, a, d) (where b_i means that all agents in $N\backslash i$ become victims). It is also the solution to the maximization problem in the solidarity upper bounds property. Note that if there exist $j \in Ri$ and $k \in Sj$ with $k \neq i$, then $e_i^{0i} > e_i^*$: agent i could pollute more in the absence of k. Doing so, under the PP regulation mechanism t^{PP}, he could enjoy a welfare of $b_i(e_i^{0-i}) - d_i(p_i^{0-i}) + t_i^{PP}(e^{0-i}) = \max_{e_i\geq 0}[b_i(e_i) - \sum_{j\in Ri} d_j(a_{ij}e_i)]$, which is higher than his welfare under the emission plan

e^*. In Moulin's terms, the presence of other polluters exhibits a negative group externality on polluter i. Solidarity upper bounds requires that every polluter who creates this negative group externality should take up a share of it. For victim only polluters $i \in V$, the solidarity upper bound is equal to zero which is their welfare under the PP mechanism.

We now provide our characterization of the PP principle generalized to increasing marginal damages.

Theorem 2 *The polluter-pays distribution rule is the unique distribution rule that satisfies non-negativity, responsibility for pollution impact and the solidarity upper bounds.*

Proof. Let ϕ be a distribution rule satisfying non-negativity, responsibility for pollution impact and the solidarity upper bounds. Let a be a problem, $\phi(a) = z$ and $\phi^{PP}(a) = y^*$. Suppose that $z \neq y^*$. Note that for all $i \in V$, by non-negativity and solidarity upper bounds, $z_i = 0 = y_i^*$. Thus, there exists $i \in N \backslash V$ such that $z_i \neq y_i^*$.

Let $a'_{ii} > a_{ii}$. Consider the problem where a_{ii} changes to a'_{ii} and everything else remains identical, i.e. $a' = (a_{-ii}, a'_{ii})$. Let $\phi(a') = z'$, $\phi^{PP}(a') = y'^*$, and e'^* denote the efficient emission plan for a'.

By RPI,

$$z_i - z'_i = W(a) - W(a') = y_i^* - y_i'^*.$$

Now we may take limits, i.e.

$$\begin{aligned} \lim_{a'_{ii} \to +\infty} z_i - z'_i &= \lim_{a'_{ii} \to +\infty} W(a) - W(a') \\ &= \lim_{a'_{ii} \to +\infty} y_i^* - y_i'^*, \end{aligned}$$

and we obtain

$$\begin{aligned} z_i - \lim_{a'_{ii} \to +\infty} z'_i &= W(a) - \lim_{a'_{ii} \to +\infty} W(a') \\ &= y_i^* - \lim_{a'_{ii} \to +\infty} y_i'^*. \end{aligned}$$

Note that $\lim_{a'_{ii} \to +\infty} \max_{e_i \geq 0} [b_i(e_i) - \sum_{j \in Ri} d_j(a'_{ij} e_i)] = 0$. Therefore, by non-negativity and solidarity upper bounds, both $\lim_{a'_{ii} \to +\infty} z'_i = 0 = \lim_{a'_{ii} \to +\infty} y_i'^*$. But now we obtain

$$z_i = W(a) - \lim_{a'_{ii} \to +\infty} W(a') = y_i^*,$$

which contradicts $z_i \neq y_i^*$.

Second, we show that ϕ^{PP} satisfies RPI, non-negativity and solidarity upper bounds. From (17) it is straightforward that ϕ^{PP} satisfies RPI because e^{0i} is optimal for both (N, b_{-i}, a, d) and (N, b_{-i}, a', d) whenever i's pollution impact is reduced (with $a'_{lj} = a_{lj}$ for all $l \in N\backslash\{i\}$ and all $j \in N$). Since e^{0i} can be implemented as an emissions plan in the problem (N, b, a, d), $W(a) \geq \sum_{j\in N}(b_j(e_j^{0i}) - d_j(p_j^{0i}))$ and, therefore, by (17), ϕ^{PP} satisfies non-negativity. For solidarity upper bounds, first note that by convexity of d_j,[11] we have for any $e_i \geq 0$,

$$d_j(a_{ij}e_i) \leq d_j(a_{ij}e_i + p_j^{*-i}) - d_j(p_j^{*-i}),$$

where $p_j^{*-i} = \sum_{k\in Rj\backslash i} a_{kj}e_k^*$. Therefore, for any $i \in N$ and $e_i \geq 0$,

$$b_i(e_i) - \sum_{j\in Ri}\Big(d_j(a_{ij}e_i + p_j^{*-i}) - d_j(p_j^{*_i})\Big) \leq b_i(e_i) - \sum_{j\in Ri} d_j(a_{ij}e_i).$$

Maximizing both sides of the inequality with respect to e_i leads to:

$$b_i(e_i^*) - \sum_{j\in Ri}\big(d_j(p_j^*) - d_j(p_j^* - a_{ij}e_i^*)\big) \leq \max_{e_i\geq 0}[b_i(e_i) - \sum_{j\in Ri} d_j(a_{ij}e_i)]. \tag{21}$$

Second, since e^{0i} maximizes $-d_i(p_i) + \sum_{j\in N\backslash i}(b_j(e_j) - d_j(p_j))$ while $(0, e_{-i}^*)$ is a possible emission plan for (N, b_{-i}, a, d), it yields a higher total welfare:

$$-d_i(p_i^{0i}) + \sum_{j\in N\backslash i}(b_j(e_j^{0i}) - d_j(p_j^{0i})) \geq -d_i(p_i^* - a_{ii}e_i^*) + \sum_{j\in N\backslash i}(b_j(e_j^*) - d_j(p_j^* - a_{ij}e_i^*)).$$

Multiplying both sides by -1, adding $W(a)$ to both sides, and using the definition of ϕ^{PP} in (17) yields:

$$\phi_i^{PP}(a) \leq b_i(e_i^*) - \sum_{j\in Ri}(d_j(p_j^*) - d_j(p_j^* - a_{ij}e_i^*)).$$

The last inequality combined with (21) shows that solidarity upper bounds holds for all $i \in N$. □

7 Conclusion: PP versus VCG

We conclude by comparing the PP mechanism with the Vickrey-Clark-Groves (VCG) mechanism applied to economies with externalities. A VCG mechanism would make each agent pay

[11] Note that $d_j(0) = 0$ and therefore, d_j is superadditive: $d_j(u) + d_j(v) \leq d_j(u + v)$ for any $u, v \in \mathbb{R}_+$.

or receive his impact on total welfare. Let e^{-i*} denote the efficient emission plan that maximizes the total welfare *without* i defined by $\sum_{j\neq i}(b_j(e_j) - d_j(p_j^{-i}))$ with $p_j^{-i} = \sum_{l\in Sj\setminus i} a_{lj}e_l$. The VCG mechanism t^{VCG} assigns to every $i \in N$:

$$t_i^{VCG}(e) = \sum_{j\neq i}(b_j(e_j) - d_j(p_j)) - \sum_{j\neq i}(b_j(e_j^{-i*}) - d_j(p_j^{-i*})).$$

Under the VCG mechanism, each agent i obtains the total welfare net of the welfare without i at the first-best. Therefore, agent i internalizes the impact of his emission on society which means that the VCG mechanism is efficient, i.e. $\mathcal{N}(t^{VCG}, a) = \{e^*\}$. It leads to the VCG-distribution rule ϕ^{VCG} defined for every $i \in N$ by:

$$\phi_i^{VCG}(a) = W(a) - \sum_{j\neq i}(b_j(e_j^{-i*}) - d_j(p_j^{-i*})).$$

Each agent $i \in N$ obtains the difference between the welfare with and without him at the first-best. In case of unilateral externalities, for a victim-only agent $i \in V$, $\phi_i^{VCG}(a) < 0$ because i's presence in the economy reduces total welfare. Therefore, non-negativity of the distribution rule induced by t^{VCG} is violated. Indeed, agent i does not only bring his damage d_i to the economy which diminishes welfare, it also forces polluters $j \in Si$ to reduce their emissions. Hence, in addition to not being compensated for the damage $d_i(p_i^*)$, a victim i has to pay for the loss of welfare which his presence causes to the polluters, namely $\sum_{j\in Si}[(b_j(e_j^{-i*}) - d_j(p_j^{-i*})) - (b_j^*(e^*) - d_j(p_j^*))]$. For a polluter-only agent $i \in P$, the PP and VCG welfare coincide because $e_j^{0i} = e_j^{-i*}$ for every $j \in N\setminus i$ while $d_i = 0$ for any $i \in P$. Therefore $\phi_i^{PP}(a) = \phi_i^{VCG}(a)$ for any $i \in P$. With multilateral externalities pollution is a public bad and the pollution problem is closer to the public good provision framework in which the Clark-Groves mechanism has been put forward. Although the VCG mechanism satisfies responsibility for pollution impact (RPI) and the solidarity upper bounds, it fails to satisfy non-negativity. An agent i adds both new emission e_i and new damage d_i to the welfare. Agent i pollutes other agents and forces in addition them to reduce their own emissions from e_j^{-i*} to e_j^* for every $j \in Si$ and $j \in Sk\setminus i$ where $k \in Ri$ for convex damage function d_i. Therefore, under multilateral externalities, we may have $t_i^{VCG}(e^*) < 0$. It is easy to found examples in which the negative impact of his presence $t_i^{VCG}(e^*)$ to society is not compensated by the i's net benefit $b_i(e_i^*) - d_i(p_i^*)$ at the first-best so that $\phi_i^{VCG}(a) < 0$ for every $i \in N$. Under the PP principle, agents pay for the

negative impact of their emissions on society not their damage. They are indeed compensated for that.

There are important differences between the VCG and the PP mechanism. Although both implement the efficient allocation of pollution emissions as a unique Nash equilibrium, only the PP principle distributes the welfare to satisfy responsibility for pollution impact and non-negativity (for constant marginal damages) and solidarity upper bounds (for convex damages). Similarly, the mechanisms proposed by Duggan and Roberts (2002) and Montero (2008) allocate pollution emissions efficiently. The advantage of the PP mechanism is that, in addition to achieving efficiency, it distributes the welfare fairly in the sense of the above three requirements. Our results strongly support the use of our interpretation of the PP principle in pollution environments.

References

Ambec, S., and Y. Sprumont (2002): "Sharing a River," *Journal of Economic Theory* 107, 453–462.

Ambec, S., and L. Ehlers (2008): "Sharing a River among Satiable Agents," *Games and Economic Behavior* 64, 35–50.

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

Carraro, C. and D. Siniscalco (1993) "Strategies for the International Protection of the Environment," *Journal of Public Economics* 52, 309–328.

Chander P. and H. Tulkens (1997): "The Core of an Economy with Multilateral Environmental Externalities," *International Journal of Game Theory* 26, 379–401.

Duggan, J. and J. Robert (2002): "Implementing the Efficient Allocation of Pollution," *American Economic Review* 92, 1070–1078.

Finus, M. and S. Tjøtta (2003):"The Oslo Protocol on sulfur reduction: the great leap forward?," *Journal of Public Economics* 87, 2031–2048.

Fleurbaey, M. (2008): *Fairness, responsibility and welfare*, Oxford University Press, Oxford.

Kolstad, C. (2000): *Environmental Economics*, Oxford University Press, Oxford.

Mäler, K.-G. and A. de Zeeuw: (1998) "The acid differential game," *Environmental and Resource Economics* 12, 167–184.

Montero, J.-P. (2008): "A Simple Auction Mechanism for the Optimal Allocation of the Commons," *American Economic Review* 98, 496–518.

Montgomery, W.D. (1972): "Markets in Licenses and Efficient Pollution Control Programs," *Journal of Economic Theory* 5, 395–418.

Moulin, H. (1990): "Uniform externalities, two axioms for fair allocation." *Journal of Public Economics* 43, 305–326.

Ostrom, E. (1990): Governing the commons. Cambridge University Press, Cambridge.