A Model of Endogenous Institutions*

James A. Robinson† **Alexander Vostroknutov**‡

Ekaterina Vostroknutova§

June 3, 2026

Abstract

We propose a new game-theoretic framework to model institutions and norms that guide behavior of agents within them. We use this framework to show how institutions can emerge from the Hobbesian "state of nature" and how they evolve due to shifts in economic environment. To the standard economic agent who maximizes consumption utility we add another motivation, the desire to cooperate with others by following social norms, and show that such moral agent has economic incentives to cooperate with strangers. When strangers cannot be trusted, institutions for facilitation emerge where a facilitator collects payment for increasing trust among moral agents who want to enter in a business relationship. We show how this general mechanism can encapsulate any game with observable actions as the model of production process. We use the model to define inclusive and extractive institutions and to show how the norms within these institutions can be influenced by the parameters of the game. Unlike the purely selfish agents, moral agents have their own perception of right and wrong and are able to self-organize. This makes inclusivity of institutions the main goal of the welfare analysis. Promoting cooperation emerges as the new normative principle for economic policy.

Topics: *Institutions and Growth, Social Norms and Poverty, Economic Growth Policy.*
Keywords: *endogenous institutions, social norms, inclusive/extractive institutions, growth, institutional policy.*
JEL Classification: *E61, E71, H20, O11, O17, O43.*

*All mistakes are our own. The findings expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank or its member countries. We would like to thank Milan Brahmbhatt, Frederico Gil Sander, Norman Loayza, Stephen Ndegwa, Hannes Rusch, Gaute Solheim and the seminar participants in IMT Lucca, DECIG Seminar Series at the World Bank, and MBEES and MLSE at Maastricht University for valuable comments.

†Harris School of Public Policy and Department of Political Science, University of Chicago.
e-mail: jamesrobinson@uchicago.edu.

‡Department of Microeconomics and Public Economics (MPE), Maastricht University, The Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl.

§The World Bank. e-mail: evostroknutova@worldbank.org. Corresponding author.

1 Introduction

The importance of institutions for economic growth and development is well-known [Acemoglu, 2005, Acemoglu et al., 2005, 2019, North, 2008]. Recent evidence on the historical emergence and development of institutions suggests that some institutions can spur economic growth while others can lead to stagnation [Acemoglu and Robinson, 2020, further AR]. AR also note that building institutions conducive to economic growth requires an enormous and sustained effort from governments and societies. Such *inclusive* institutions are located in the "narrow corridor" between a fully disorganized "state of nature," or Warre, and an all-controlling Leviathan of a powerful government. AR provide evidence of institutional design and historical path-dependence that have led countries to or away from this narrow corridor.

Inclusive institutions possess features that make them effective in creating economic growth, however only a handful of developing countries have been able to engineer such institutions and convergence to the narrow corridor. The majority of developing countries suffer through *extractive* institutions, where a small proportion of the population enjoys the benefits that all workers of the economy create, and are not therefore conducive to long-term economic growth per capita. This is partly because our understanding of the exact mechanisms by which the effects of institutions on growth are transmitted is not universal. The lack of models of institutions is especially felt in policy applications, where the ability to precisely predict institutional change would lead to more effective policies to spur economic growth, especially in developing countries [Lopez-Calva et al., 2017, Lall et al., 2024].

In this paper, we study one mechanism of how institutions arise from cooperative motivations of economic agents and from their interactions with the economic environment. In applying game-theoretic modeling to account for the emergence and evolution of institutions, we complement the existing literature where similar approaches are used to study the evolution of social norms that drive cooperation [e.g., Young, 1998, Acemoglu and Jackson, 2015].

Indeed, the general consensus across many literatures is that institutions arise to support cooperation of economic agents [Nee, 1998, Ostrom, 2000, Keefer and Knack, 2008, Henrich, 2015a]. Cooperation is a driving force that creates and sustains societies that are resilient to warfare, economic or climate shocks, and other challenges. For example, Henrich [2015a] provides evidence from historical and current indigenous populations in Alaska, Australia, and Papua New Guinea of the crucial role that cooperation played in the ability of these small-scale societies to not only survive in challenging environments but also to thrive and proliferate. In modern, complex societies, cooperation between the state and the people is also one of the main features of inclusive institutions. As AR note, "in the corridor [of inclusive institutions] the state and society do not just compete, they also cooperate. This cooperation engenders greater capacity for the state to deliver the things that society wants and foments greater societal mobilization to monitor this capacity."

To better account for the specific types of cooperation that arise in institutional settings, the connection between norm-abiding behavior and business activity that has been pointed out by Arrow [1972]. We introduce one small but fundamental feature in our framework that is different from the previous

literature. Instead of a *single motivation*, the desire to have more consumption utility, the agents in our model trade off *two motivations:* the standard preference to have more consumption and *the desire to cooperate with others*. This second motivation allows us to model formation of institutions and to predict how they might change in various circumstances.

While the introduction of the motivation to cooperate is novel in economics, large swaths of literatures in other social sciences have converged on the view that humans are a social species that has evolved special mental capacity for cooperation. The reason this capacity has evolved is that cooperation is more profitable than no cooperation [e.g., Bicchieri, 2006, Henrich, 2015a, Fehr and Schurtenberger, 2018, Laland, 2018]. The innate preference for cooperation manifests itself through the desire to adhere to *social norms*, or common beliefs about how things ought to be done. In economics, this motivation was first introduced by Kessler and Leider [2012] as an additional term in the *norm-dependent utility function*. Norm-following or *moral* agents maximize the sum of their consumption utility and a measure of social appropriateness defined on outcomes (the agents prefer more socially appropriate outcomes to less socially appropriate ones).[1]

The introduction of the cooperative motivation dramatically changes the way agents behave and helps understand how institutions emerge. In our framework, moral agents care about the social appropriateness of their actions, which tends to be the highest when they successfully cooperate with each other. This is the motivational force that pushes them to find ways to create new cooperative enterprises, and leads to the formation of institutions. However, selfish motivation still plays a role and this implies that moral agents might not be able to cooperate (even if they want to) due to lack of *trust*. Indeed, trust to strangers is typically not high, even in developed countries (one wouldn't enter a long-term business relationship with a random person they meet on the street).

To resolve the problem of low trust, people use institutions and norms that facilitate economic and social transactions. In addition to groups with already high trust (such as friends and family, or common acquaintances), Henrich [2015a] summarizes evidence from indigenous populations that kinship, marriage, and ritual institutions involve social norms that tightly bind males from different groups in interdependent social webs. These norms, rituals, and status figures in charge together facilitate cooperation in the environment of low trust. In our model, we incorporate these elements by introducing an additional agent, the *facilitator,* whom all moral agents trust in a given context. The resulting *institution for facilitation* maintains high trust among moral agents in the context of the enterprise that they have in mind and defines the payments that the agents should make to the facilitator. The strategic interactions between the facilitator and the agents that involve choices of cooperative efforts and payments to the facilitator eventually determine how inclusive or extractive the institution for facilitation is.

The paper is structured as follows. In Section 2 we describe the general approach and the new framework to study the emergence of social norms by Kimbrough and Vostroknutov [2022], further KV. We

[1] Kessler and Leider [2012] proposed the general idea of mixing consumption and normative considerations in one utility function. They did not aim at specifying exactly how the norm function is constructed. Kimbrough and Vostroknutov [2022] took it further and proposed a theory of injunctive norms that specifies how norms emerge from the context of payoffs in the game.

review the literature and show how this framework allows us to construct complex institutions [also termed "packages of social norms" Henrich, 2015b] from the building blocks of social norms that emerge in individual interactions.

In Section 3, we present the general game-theoretic setup and use the illustrative example of the Public Goods game to analyze the behavior of moral agents. This game serves as an abstract representation of a cooperative enterprise, a "firm," that can emerge or not depending on whether agents decide to cooperate. We show similar results for general games with observable actions and also relate the parameters of the model to the intuitive notion of trust that is considered by many authors to be the necessary ingredient of economic development [Lederman et al., 2002, Keefer and Scartascini, 2022].

In Section 4, we define the institution for facilitation and show how the norm (the amount of payment made to the facilitator) emerges endogenously from the parameters of the game and trust weights.

In Section 5, we propose a game structure that describes the relationships between the facilitator and the agents. The analysis of equilibrium behavior in the game with facilitator then paves the way to natural definitions of *inclusive* (or norm-abiding) and *extractive* (or norm-breaking) institutions from AR. These definitions play the key role in the welfare analysis that comes out of our framework. Specifically, in the world where agents themselves determine their own morality and self-organize, the goal of economic policy also changes compared to the case of selfish agents where only the social planner knows what is socially desirable for them. In the world of moral agents, the focus of economic policy is to provide opportunities for cooperation, rather than redistributing resources according to the morality of the social planner that is external to the agents. In other words, *the goal of economic policy becomes making institutions more inclusive and conducive to cooperation.* Such policies should help moral agents to create a better, more fair redistributive system that fits their own moral views.

Finally, Section 6 concludes with a discussion of further directions of research and the implications of our framework for economic and institutional policy.

2 Literature and Methodology

According to the long tradition of institutional economics, institutions can be good or bad for economic development [North, 2008]. But while Western-style economic and political institutions have been shown to correlate with economic growth, the literature also documents many failed attempts at imposing them in developing countries [Putnam et al., 1992]. The policy dilemma of the century that can be succinctly formulated as "How to change institutions?" has attracted a lot of attention in the literature, but still remains unsolved.

Empirical analysis of the importance of institutions for development has traditionally treated institutions as exogenous variables in cross-country regressions, with proxies such as rule of law, property rights protection, democracy index, etc. The literature analyzing specific institutions, such as legal regulations guiding financial market functioning, is mostly not generalizable to other situations or contexts. Similarly to Shirley [2005] and Knack and Keefer [1995], Rodrik et al. [2002] note that empirical studies

show significant regularities in how institutional variables tend to dominate others in explaining growth and social progress, "but these studies lack a theory that would transform regularities into causal explanations."

Nevertheless, important theories of institutions have been proposed in the literature. Ostrom [2000] uses social dilemmas to reason about the nature of collective action and social norms in Common Pool Resource problems [see also Kimbrough and Vostroknutov, 2015]. Hart [1989] and Holmström and Roberts [1998] provide an economic perspective on the boundaries of a firm. Acemoglu and Jackson [2015] study the emergence of a general institution for "cooperation." These models capture certain elements of institutional influence on economic behavior, but are either too general or too specific to be used as a workhorse model for the wide variety of institutions encountered in applied work. In such a model, institutions would need to enter *endogenously* [see also Engerman and Sokoloff, 2005], so that it is possible to consider the emergence and dynamic development of institutions depending on a set of parameters. This is needed for testing the model in practice as well as for policy implications and can only be achieved if institutions are introduced at the micro level, as social norms that aggregate into institutional packages [Henrich, 2015b].

In this paper, we offer a model with these features. An important innovation is the introduction of moral agents, who compute normative values (social appropriateness of outcomes) from the specific context of the institution (payoffs in the game). This allows us to pin down exactly the utility-maximizing behavior of moral agents within the institution and how it changes with the parameters.

It may seem that normative values in a given institution might not be determined explicitly by the current payoffs (as is proposed in KV) but might also be influenced by the past history of the institution, local beliefs, traditions, and other specific cultural features. This is indeed true for many *existing* institutions and has been studied extensively [e.g., Bicchieri, 2006, 2016]. However in the specific context of *emerging* institutions, past history may not play the dominant role. For example, new business relationships may arise with the new opportunities available to the agents at the given moment in time. Thus, we use the model of injunctive norms of KV where normative values are computed from the payoffs that represent current business opportunities. As was suggested by Henrich [2015a], new norms that emerge in this way later crystallize into the customary rules and regulations of the institution.

KV describe a specific case of norm formation in any context that is based on the idea of aggregating *dissatisfaction*. Specifically, it is assumed that the social appropriateness of each given allocation is inversely proportional to the summed dissatisfactions of all agents in that allocation. The dissatisfaction of each agent in a given allocation is high when there are many other counterfactual allocations in the context that give the agent more consumption utility, and low otherwise. Thus, dissatisfactions express personal grievances of agents in a given allocation, and the normative value of the allocation is formed by aggregating all these dissatisfactions. In other words, the allocation is considered more socially appropriate, the less aggregated dissatisfaction it evokes.[2] We will refer to agents who care about

[2]Experiments by Merguei et al. [2022] and Panizza et al. [2021] perform direct and rather stringent tests of KV's model and find a good fit with the data. Tests on existing experimental datasets in Kimbrough and Vostroknutov [2022] have shown that

dissatisfaction of others in the way described by KV as *moral agents*.

Following KV, we also introduce social weights (or trust weights) that measure the importance of dissatisfaction of each moral agent in the aggregation.[3] These weights provide an important mechanism to account for social behavior within and without an in-group [Chen and Li, 2009], and translate intuitively to our definition of trust.

The above features of the model help us with the main task of this paper: to model not only interactions of agents with the existing social norms but also to allow for norms to arise endogenously depending on the existing situation. In what follows, we use the properties of injunctive norms of KV to model the transition from extractive to inclusive institutions in the AR framework.

3 Model

3.1 General Formulation

The main purpose of this paper is to introduce a new modeling framework based on norm-dependent utility and to show how it can be used to model institutions, their emergence and change, and how to assess properties of such institutions, specifically how inclusive or extractive they are. In the framework, the production process within the institution can be general and modeled as virtually any game. Making some light assumptions described below, we consider such games as embedded in a larger strategic interaction with a facilitator. Then we formulate some general results pertaining to the equilibrium behavior in such extended games—or institutions—and show how they can be classified as inclusive and extractive.

We start with the general description of games to which our arguments and definitions can be applied. To specify the *institutional problem* that moral agents face, we assume that there are N agents who would like to enter some business relationship with the purpose of jointly producing something. The process of interaction among agents (necessary for production) is described by some finite extensive-form game Γ with game outcomes in some (bounded) set $C \subseteq \mathbb{R}^N$ representing possible payoffs that agents may receive. For example, C may include payoff vectors where agents do manage to successfully produce something and all get richer, or outcomes where agents fail to work together to the fullest degree and only have partial success (or anything else). Thus, we consider some game (Γ, C), where Γ describes the actions of the players and the histories of actions that lead to outcomes in C. Next, we provide some structure and several (not too demanding) restrictions on Γ that allow us to build our arguments.

Trust weights. Since we deal with moral agents who maximize norm-dependent utility, we need to add

the model of dissatisfaction-based norms can account for several important puzzles in social behavior. For example, radical context-dependence in Dictator games [List, 2007, Bardsley, 2008] or extreme switches of moral principles with small changes in payoffs in bargaining situations [Galeotti et al., 2018]. See Vostroknutov [2020] for discussion.

[3] In each allocation, the dissatisfaction is aggregated by summing up individual dissatisfactions of all agents. The weights multiplying these individual dissatisfactions act as measures of trust to individual agents. When an agent is not trusted (low weight), her dissatisfaction will count little in the aggregation, so that it may become socially appropriate to give her less than others (and vice versa for the agents with high trust weight).

a set of additional parameters to the description of Γ. Specifically, we assume that there are some trust weights $\tau_{ij} \in \mathbb{R}$ defined for each pair of players $i, j \in N$ (how much agent i trusts/cares about/takes into account agent j). These weights are used by moral agents to compute social appropriateness of various outcomes in the game (the higher the weight, the more "deserving" of higher payoff the agent becomes). As also discussed in KV, we use simplifying assumptions that for all $i, j \in N$ we have $\tau_{ii} = 1$ (i cares about herself to degree 1); $\tau_{ij} = \tau_{ji}$ (two moral agents always trust each other to the same degree); and all trust weights are common knowledge among all players in N. Below in the "Perfect Information" paragraph, we discuss what happens when we relax these assumptions.

Endowments and scalability. As we mentioned in the Introduction, we propose that institutions may arise due to the process of facilitation of relationships among moral agents by a facilitator. This is needed when moral agents do not trust each other enough to successfully work together within (Γ, C), so another player called facilitator is added to the game. We assume that agents in N can pay facilitator for her ability to increase trust among them (so that with facilitation they can work together better). Thus, to accommodate this mechanism we assume that each agent has some fixed and equal endowment that can be used to pay the facilitator. For tractability, we assume that all players pay the facilitator the same fee and are left with the amounts $a > 0$ each to play the game. We consider (Γ, C) as *scalable* in the following simple sense. We assume that $C = C(1)$ describes the payoffs for players in case when $a = 1$, in other words, the endowments used in the game are all equal to 1 for all players. We assume that the game $(\Gamma, C(a))$ for arbitrary endowment $a > 0$ can be obtained from (Γ, C) by simply multiplying all payoffs of all agents in all outcomes in C by a. Notice that the description of moves in Γ does not change, so we assume that the production process happens by the same rules. This device allows us to think in a simple way about the consequences of playing $(\Gamma, C(a))$ after different payments to the facilitator. If agents give away a lot of their endowments, they will be left with little resources (low a) and will not be able to reap high benefits from cooperation. So, payments to facilitator decrease the potential benefits of working together.

Social dilemma structure. We mentioned that moral agents try to resolve the institutional problem that they face (lack of trust) and that is why they require the facilitator. An important part of this construction is the implicit assumption that agents cannot achieve the desired production levels within (Γ, C) without facilitation when they do not trust each other. In other words, the incentives that drive standard, selfish agents in (Γ, C) (who do not take norms into account) do not lead to outcomes that are favorable for all agents. This idea can be expressed in two simple assumptions. First, that the set of payoff vectors C contains something that is better for all players than their original endowment of 1. In other words, there is some vector $c \in C$ with $c_i > 1$ for all $i \in N$. This reflects the general idea that agents who play (Γ, C) can achieve *growth* when they work together – there are ways to increase everyone's well-being. Second, we assume that none of such vectors c that lie on the Pareto frontier of C are Nash equilibria of (Γ, C) when played by selfish players. This is the typical structure of a social dilemma where selfish players cannot achieve cooperative outcomes because these outcomes are

not sustainable in equilibrium. We hold an opinion that any purposeful interaction of people necessarily involves some elements of cooperation and coordination and that such social dilemma structure not only reflects the nature of real production processes, but also suggests that successful production and growth are impossible without sustainable cooperation.

Unique norm. Another assumption that we make for analytical tractability is that the game (Γ, C) is such that the injunctive norm function $\eta_i(c)$ computed for agent i at some outcome $c \in C$ has a unique maximum (for all $i \in N$). We call such maximum the *norm* for player i. Following KV, we provide the exact analytical formula for $\eta_i(c)$ and it turns out that this assumption is satisfied for all generic games. It takes, in fact, some imagination to come up with games where there is no unique norm. This typically happens when games have symmetric but unequal payoffs like in the Battle of the Sexes, or in tournaments, games with very specific structure (see KV). When C is a convex N-dimensional polytope (this class of games includes most games with continuous action spaces studied by economists, like Public Goods or Trust games), KV show that η_i is a piece-wise concave quadratic form, which guarantees unique maximum in generic cases.

Perfect information. Our final assumption is that (Γ, C) does not contain any informational asymmetries. Specifically, we assume that (Γ, C) is a game with observable actions, where in each period players participate in a normal-form game and the outcome of their play becomes known to everyone before the next period (where some other normal form can be played). This assumption rules out any information asymmetries with regard to what happens in (Γ, C) and is necessary to make sure that the normatively best outcome can be achieved if players want to do that. It is easy to imagine how this may not be so with asymmetric information. For example, players might have different opinions about which sequence of actions leads to some normatively good outcome $c \in C$. This difference may prevent them from reaching it. Similarly, we assumed above that players have common knowledge of trust weights. This assumption boils down to the same effect: without it we cannot guarantee that players can achieve the normatively best outcome, which is important for our arguments below.

To summarize, we consider scalable game with observable actions (Γ, C) that has social dilemma structure and unique norm-maximizing outcomes for each player. For the group of N moral agents, achieving outcomes on the Pareto frontier can be impossible if they do not trust each other sufficiently, which is also true for selfish agents (there are no Nash equilibria on the Pareto frontier). Thus, we assume that no growth or production are possible without some form of cooperation. This constitutes agents' institutional problem. In a world with low trust, facilitation of trust relationships will be conducive to the emergence of cooperation. If moral agents find a paid agent they can all trust, a facilitator that improves trust between them, their desire to adhere to norms will also result in cooperative behavior, production, and growth.

3.2 Example of Public Goods Game

To better understand the general result, we also consider an example where the production process is described by the Public Goods game (PG). This game succinctly represents all important assumptions about the production process, is representative of the general results, and presents them in a stylized setting. This example thus aims to provide intuition for the new mathematical concepts and demonstrate how game-theoretic arguments can be constructed in other games.

We assume that two moral agents desire to profit from mutual cooperation in the Public Goods game (PG), which is similar to game forms used by Ostrom [1990] to study common pool resource problems. This interaction can be seen as representing various forms of "doing business," which can produce surplus and consequently growth of wealth through cooperation (or not). This social dilemma structure reflects the various views in institutional economics, where institutions facilitate interactions of economic agents and impact transaction costs [North, 2016]. Given that any form of economic activity requires a certain degree of trust and cooperation/coordination among the people involved, and similar to Ostrom [1990], we focus on a specific form of transaction costs, those related to trust and cooperation. We consider PG as a versatile example of a social dilemma that embodies potentially variable levels of productivity (through the PG multiplier).

To describe how PG is played, suppose that before the game each player $i \in \{1, 2\}$ has some endowment $w \geq 0$.[4] In the game, player i chooses an amount to contribute to the public good with some PG multiplier $p \in [\frac{1}{2}, 1)$. Suppose that player i contributes $x_i \in [0, w]$, then her wealth after the game is

$$w_i = w - x_i + p(x_i + x_{-i}). \tag{1}$$

In other words, player i keeps the part of the endowment that was not contributed ($w - x_i$) and gains the return from the public good (the sum of contributions $x_i + x_{-i}$, where x_{-i} stands for the contribution of the other player, times p). Notice that the multiplier p can be thought of as a measure of *productivity* and defines how profitable the cooperation is. It can be influenced by various factors (e.g., it is high when good-quality public services are provided and low otherwise).

When agents are standard consumption utility maximizers, the unique Nash equilibrium of this game is to contribute nothing ($x_i^* = 0$ for all i). This represents the classic case of underprovision of public goods and also satisfies our assumption that there are no Nash equilibria on the Pareto frontier. When player i is assumed to be a moral agent, she chooses x_i to maximize her *norm-dependent utility*

$$U_i(x_i; x_{-i}) = w_i + \phi_i \eta_i(w_i, w_{-i}). \tag{2}$$

The utility function U_i consists of two terms: the standard linear consumption utility w_i and the normative term $\phi_i \eta_i(w_i, w_{-i})$ representing the desire to follow norms.[5] Here we implicitly see w_i as the

[4] A generalization to PG with unequal endowments is very interesting and easy to implement theoretically. However, for the sake of expositional simplicity we choose to not introduce it in this paper and leave it for future research.

[5] The idea to explicitly have norms in the utility function was first introduced by Kessler and Leider [2012] and then was

function of x_i and x_{-i} as in (1); the coefficient $\phi_i \geq 0$ defines i's propensity to follow norms; and $\eta_i : C \to \mathbb{R}$ is the *norm function of player* i that defines the *social appropriateness* of each possible outcome (w_1, w_2) in the game and arises endogenously from the set of all achievable wealth allocations defined by $C = \{(w_1, w_2) \mid (x_1, x_2) \in [0, w]^2\}$. Specifically, following KV we define η_i as

$$\begin{aligned}\eta_i(w_i, w_{-i}) = \eta_i(w_i, w_{-i}; \tau) &= -[D_i(w_i) + \tau D_{-i}(w_{-i})] \\ &= -\left[\int_{c \in C} \max\{c_i - w_i, 0\} dc + \tau \int_{c \in C} \max\{c_{-i} - w_{-i}, 0\} dc\right]. \quad (3)\end{aligned}$$

Here, the notation $\eta_i(w_i, w_{-i}; \tau)$ emphasizes that the social appropriateness expressed by the norm function depends on $\tau \geq 0$, the *social weight* that player i attaches to the other player $(-i)$. $D_i(w_i)$ stands for the *personal dissatisfaction* that player i feels should she receive wealth w_i after the game. This personal dissatisfaction, in its turn, is the sum of dissatisfactions due to all possible other allocations in C where player i receives more consumption utility than w_i (so, i is dissatisfied because she could have had more consumption). This is expressed with the max operator in the second line of (3), that measures dissatisfaction at w_i due to some other possible consumption utility c_i (we notate $c = (c_i, c_{-i})$ since C is a subset of $\mathbb{R}^2$).

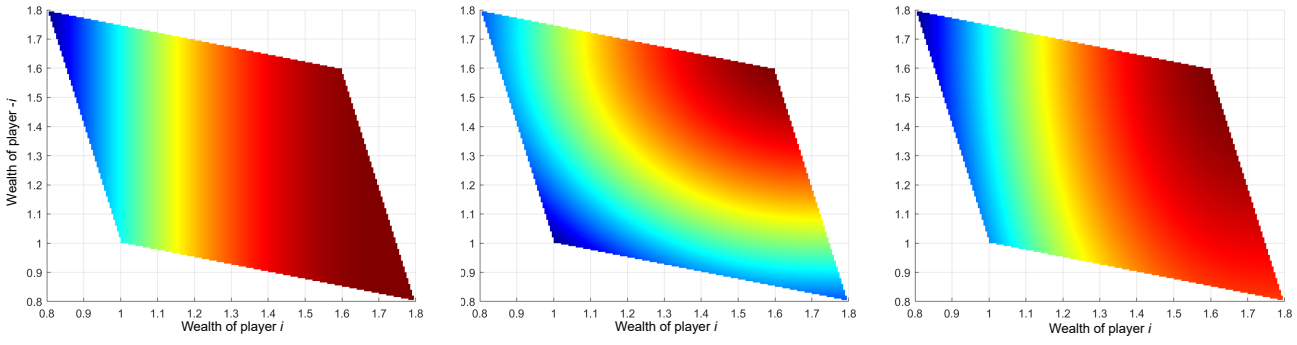

Figure 1: **Left.** The set of allocations C with $w = 1$, $p = 0.8$, and $\tau = 0$ (dark blue - the lowest social appropriateness; dark red - the highest social appropriateness). **Middle.** Same as the left panel only with $\tau = 1$. **Right.** Same as the left panel only with $\tau = 0.2$.

Thus, the social appropriateness $\eta_i(w_i, w_{-i})$ of outcome (w_i, w_{-i}) from the perspective of player i is the negative of the weighted sum of the personal dissatisfactions of the two players ($D_i(w_i)$ and $D_{-i}(w_{-i})$) with the latter weighted by the exogenously given social weight τ. The idea here is that player i feels that outcomes in C are more socially appropriate when they evoke less dissatisfaction from both players, though the dissatisfaction of the other player $(-i)$ can be inflated or discounted in comparison to i's own dissatisfaction (which has weight 1). For example, $\tau = 0$ would imply that player i does not normatively care about player $-i$ as a human being at all (e.g., slavery). From her perspective social appropriateness only increases when her own dissatisfaction goes down (regardless of what happens to the dissatisfaction of $-i$). The left panel on Figure 1 shows such norm function. When $\tau = 1$ we get

experimentally tested by Krupka and Weber [2013].

the case when player i treats the other player equally to herself. She cares about the dissatisfaction of the other player as much as she cares about her own (e.g., family). The norm function η_i for this case is shown in the middle panel of Figure 1. When $\tau \in (0, 1)$ we get the situation where i cares somewhat about the dissatisfaction of $-i$, but not as much as about her own. This is the case, for example, when $-i$ belongs to another social group than i. Equal partners, who are not related through family ties, might have a τ above a certain threshold, but less than 1. The right panel of Figure 1 illustrates.[6]

Given these definitions, we can talk about the *norm from the perspective of player* i as an allocation $c_i^* = \arg\max_{c \in C} \eta_i(c)$ that has the highest social appropriateness. It is clear that in general, when $\tau \neq 1$, the two players will not "see eye to eye" with respect to what they consider as the norm in this game. However, when both players treat each other as equals ($\tau = 1$ for both of them), they will agree on the norm, which should make cooperation easier to achieve (in this case $c_i^* = c_{-i}^*$ correspond to the outcome where both players contribute fully to the public good, see the middle panel of Figure 1). Notice, however, that players will contribute full amounts in equilibrium even when τ is less than 1 but close enough to it. We characterize the Nash equilibria of the Public Goods game in a proposition.

Proposition 1 (PG). *When both ϕ_i are close enough to 0, there is a unique Nash equilibrium that is the same as in the PG game with selfish players (contribute nothing: $x_i^* = x_{-i}^* = 0$). The rest of the statements apply to the case when both ϕ_i are high enough. When $\tau > (1-p)/p = \tau^*$, there is a unique Nash equilibrium where both players contribute everything ($x_i^* = x_{-i}^* = w$). When $\tau < \tau^*$, there is a unique Nash equilibrium where both players contribute nothing ($x_i^* = x_{-i}^* = 0$). When $\tau = \tau^*$ any equal contribution $x_i^* = x_{-i}^*$ is Nash equilibrium.*

Idea of the Proof. We show the complete proof in Appendix A. To find the characterization of equilibria in the case of high ϕ_i, we use the result from KV (Compromise Theorem). This theorem applies to cases when C is any convex N-dimensional polytope (true for the PG game) and shows that the norm function η_i in such cases is always a piece-wise concave quadratic form. We use this fact together with some continuity properties to characterize the fixed points of the Best-Response correspondence (Nash equilibria).

3.3 Unique Cooperative Equilibrium in General Games

This proposition for PG uncovers an important property of equilibria in games played by moral agents (with high enough ϕ_i). Specifically, moral agents can support full cooperation (full contributions to the public good) even when their trust to each other is not perfect: unique cooperative NE exists for a range of $\tau \in (\tau^*, 1]$. This means that cooperation is not something that can only be reached in the limit when $\tau = 1$ and everyone cares about everyone else like for themselves (which might be unrealistic). According to this result, cooperation can be achieved even by players who are not fully trusting each other, which makes it much more plausible (we also see a lot of cooperation in PG experiments).

[6] It is important to emphasize at this point that this model has been tested against many experimental datasets (see KV) and shows a very good fit to actual human behavior.

We can generalize the important pieces of this result to all possible games with observable actions (Γ, C). To do that, let us define the norm function η_i for the general case with the set of payoff vectors C and some vector $y = (y_1, ..., y_N) \in C$ as

$$\eta_i(y) = -\left[D_i(y_i) + \sum_{k \in N \setminus i} \tau_{ik} D_k(y_k)\right]$$
$$= -\left[\int_{c \in C} \max\{c_i - y_i, 0\} dc + \sum_{k \in N \setminus i} \tau_{ik} \int_{c \in C} \max\{c_k - y_k, 0\} dc\right]. \quad (4)$$

We notate $c = (c_1, ..., c_N) \in C$. When C is discrete, integrals are replaced with sums. This is the straightforward generalization of equation (3).

For the general case we cannot show the unique cooperative NE result for only "high enough" ϕ_i. The reason is that (Γ, C) can have arbitrary payoffs, and thus we can only make the argument asymptotically. Let us assume that $\phi_i \to \infty$ for all $i \in N$ or that the agents are so moral that they only care about norm functions and disregard their own payoff. In this case we can think of their utility as being defined simply by $\eta_i(y)$. For this case we can formulate the following proposition.

Proposition 1 (General). *When all ϕ_i are close enough to 0, the SPNE of (Γ, C) are the same as when (Γ, C) is played by selfish agents. When $\phi_i \to \infty$ and $\tau_{ij} = 1$ for all $i, j \in N$ we have $\eta_i(y) = \eta(y)$ for all players i. In this case, all players will choose actions in Γ that lead to the unique most appropriate outcome $y^* = \arg\max_{y \in C} \eta(y)$. This will be the unique (cooperative) Subgame-Perfect Nash equilibrium of (Γ, C).*

Idea of the Proof. For the case $\phi_i \to \infty$, the assumptions of the proposition essentially render all strategic interactions among players unnecessary (under assumptions we made in the beginning that there is no asymmetric information, etc.). The fact that all players only care about the norm function and it is the same across them all, makes everyone's preferences over outcomes in C identical. Thus, all players will choose actions on any stages of Γ to reach the unique (by assumption) mostly appropriate outcome. The non-cooperative game structure Γ becomes irrelevant. Note as well that the assumption $\phi_i \to \infty$ is only needed when $\eta(y)$ has a zero derivative at maximum. When the derivative is not zero (the case of PG and any other game where maximum is at the vertex of C), the results will go through for some finite, high enough ϕ_i. In this case, cooperation can also be sustained for some τ_{ij} close enough to 1 as in the case with PG.

This general result demonstrates where the power of social norms (as common beliefs about what is appropriate) truly comes from. In a well-organized community of N moral agents (no information asymmetries) who deeply care about norms, cooperation can emerge regardless of the strategic complexities that the agents might face. This is because agents' strong desire to follow norms turns them into players with identical preferences and they all, as a group, strive to reach the most socially appropriate outcome in any game. When agents are not moral (low ϕ_i), their behavior will be described by some SPNE of the game with selfish players, which by assumption is not efficient (does not lie on the Pareto frontier) and

corresponds to failed cooperation.

3.4 Two Definitions of Trust in the Model

As described in the previous part, moral agents cooperate or do business more willingly with people whom they trust, so the concept of trust is important to understanding the behavior and beliefs of moral agents. A widely used definition of trust states that it is the belief that someone will not behave opportunistically, as in Keefer and Scartascini [2022]. A more general definition is that trust is a situation-specific expectation about other agent's behavior [Bauer and Freitag, 2018]. In other words, trust is context-dependent. In particular, trust can be linked to propensity to follow specific social norms. For example, agents exhibit higher trust towards people who belong to their social group [Chen and Li, 2009]. To reflect this, we add a second dimension to the commonly used definition of trust.

The first dimension of trust corresponds directly to the definition of Keefer and Scartascini [2022]. In PG, this kind of trust is defined by the individual propensities to follow norms ϕ_i and ϕ_{-i} (and players' beliefs about them). Indeed, if player i believes, for example, that $\phi_{-i} = 0$, then she is sure that the other player is selfish and only maximizes his consumption utility. She *does not trust him,* because she believes that $-i$ will disregard the norms (whatever they may be) and contribute nothing in PG, which is $-i$'s strictly dominant strategy.

For the second dimension of trust (as we define it), notice that in our model i's believing that $-i$ has high ϕ_{-i} is not enough for her to trust him. If i believes that ϕ_{-i} is high but that $-i$ does not care about her (low τ in η_{-i}), then she also will not trust $-i$. This happens not because i thinks that $-i$ is selfish, but rather because $-i$ has *different norms* that discount her, i's, importance in the eyes of $-i$. For example, the members of two warring tribes might be absolutely sure that the opponent is a highly moral norm-following individual. However, they will still not cooperate with each other (contribute in the PG) because they know that both of them do not care about the dissatisfaction of the other. They *do not trust each other due to different normative views* [see e.g., Akerlof and Kranton, 2000, Chen and Li, 2009]. This is an important case of trust that we believe is the key to understanding the formation of institutions given that most human beings do follow norms of their—sometimes highly specific or imaginary—social groups most of the time.[7]

In the rest of the paper, we will assume therefore that all players are moral agents who, by definition,

[7] Our terminology with regard to τ that we call the second dimension of trust might sound unfamiliar to some readers. The way τ enters norm-dependent utility is technically similar to models of altruism where agents directly care about the utility of other players. However, the similarity is only technical. What is important is the interpretation. In case of norm-dependent utility, τ is a parameter that determines *normative views*. In other words, τ determines whether an agent will consider some allocations involving other agents socially appropriate or not. Norms are social constructs that incentivize agents to act in the society, whereas altruism is a psychological construct that does not necessarily relate to social behavior. It can be easily imagined that there are agents who are not altruistic at all, but who follow norms nonetheless. Thus, if we consider utility representation (2) as just some utility function that involves utilities of other players, then τ can be interpreted as altruism. But when we consider it as a *norm-dependent utility* with norms that are supposed to be shared among many agents, then τ obtains a different meaning, it becomes related to whom we should trust and whom we should not trust. In the example with warring tribes, τ signifies how appropriate it is to be cooperative with the members of the other tribe and not how altruistic we should be towards them.

have some propensity to follow norms (ϕ's are above zero), but that their normative views are different due to low social weights τ put on other players. Thus, we describe the world where all agents want to cooperate by following *some* norms (e.g., those of their social group), but might fail to cooperate with strangers because they think that they come from the out-group.[8]

4 Institutions for Facilitation

Institutions are "humanly devised constraints that structure human interaction," including formal constraints such as constitutions and laws and informal constraints, such as norms, conventions and self-imposed codes of conduct [North, 1990]. Describing the incentives for "doing business," Coase argued that "if the costs of making an exchange are greater than the gains which that exchange would bring, that exchange would not take place" [Coase, 1937]. Institutions thus emerge to reduce transaction costs associated with production.

To understand the functioning of endogenous institutions in our model, we must understand their emergence and trace their development. We start from the Hobbes' primordial "state of nature," devoid of any institutions or social norms and observe the emergence of institutions to support cooperation. We eventually arrive at a model with endogenous institutions, where institutional change is possible from extractive to inclusive institutions, depending on personal characteristics of agents involved (ϕ) and their trust to each other (τ).

4.1 The State of Nature

To understand how and for what purpose institutions may emerge, we need to consider the motivation and incentives of agents in what moral philosophers call the *state of nature* or *Warre* [Hobbes, 1651]. This is a hypothetical world of unorganized individuals who do not hold any common views on how things should be done and thus act mostly in their own self-interest in the absence of any shared norms, regulations, or laws. This is the world of "war of all against all" that existed in human prehistory and can still be found today in some societies. For example, Henrich [2015a] documents the extremely high rates of violent deaths among modern and prehistoric groups of hunter-gatherers. AR describe the situation in present-day Nigeria where violence between warring groups acts as a deterrent to economic development. Leeson [2007] depicts the conditions in Somalia that are similarly close to the state of nature.

Translated to the language of our model in Section 3.2, this would correspond to the situation where moral agents, although willing to follow norms, are unable to cooperate because they do not trust anyone (high ϕ, low τ). In this case, in equilibrium no two agents will contribute anything to the public good and thus there will be no growth of wealth.[9]

[8] The case with low ϕ's is not interesting anyway, as it is approximated well by standard players with selfish preferences. In this case, we know already that no cooperation can happen at all in any social dilemma.

[9] Hobbes describes the state of nature as "war of all against all," which can be likened to negative growth. In our model,

Even though the state of nature precludes wide-spread cooperation with strangers due to generally low trust, some episodic cooperation and growth of wealth might exist among moral agents who know each other well and have high social weights τ attached to each other (e.g., family, friends). A pair of friends (or cousins) with high enough τ will contribute fully to the public good, thus benefiting themselves. However, it is also reasonable to think that after some time they will exhaust all possibilities for profitable cooperation due to the fact that they most likely live in the same area and have access to the same resources. This will make the PG multiplier p, applicable to their interaction, low and the resulting growth will be negligible.

From this we can argue that moral agents would strive to cooperate with others who have access to different resources and have potentially different skills and know-how (which makes the multiplier p in the potential PG high). However, such others will typically be strangers that cannot be trusted in the state of nature (low τ). Thus, we have a situation where moral agents *want* to cooperate because it is profitable, but are *unable to do so* due to low trust. In this case, it is reasonable to believe that they will try to find *some arrangements* that will make cooperation possible.

4.2 Facilitation

The simplest possible arrangement that can expand cooperation beyond what is possible in the state of nature is *facilitation* of cooperative relationship between two moral agents by a common friend, relative, or *facilitator*. It is possible that first facilitators emerging from the state of nature were extended family members related to agents through family ties (e.g., cousins, in-laws, or clan members). In later times and in societies that started to interact with strangers, facilitators were traders or other individuals (e.g., priests) who managed to gain trust of separate groups of people with different norms or social identities. This allowed them to act as intermediaries between them also facilitating economic relationships.

Consider the Public Goods game played by two moral agents as in Section 3.2 with equal and low social weights $\tau_2 < \tau^*$ attached to each other, which is below the threshold τ^* above which the Nash equilibrium in the PG is to fully contribute. Thus, in the PG the agents will optimally choose to contribute nothing when their mutual social weights are τ_2. Now suppose that there is a facilitator agent f with whom both agents maintain friendly relationships, so that the mutual social weights between any agent $i \in \{1, 2\}$ and f are $\tau_1 > \tau^*$. This means that agent i would cooperate with f if the opportunity arose. The left panel of Figure 2 illustrates: the dotted line emphasizes the impossibility of cooperation.

Despite the possibility of cooperation with f, the agents want to cooperate with each other as this brings more profit than cooperation with f (see the argument above). Therefore, it is possible that the three agents agree on the following deal. Agent f facilitates the relationship between agents 1 and 2 by "vouching" for each agent as a good cooperative partner (since both are his friends). In the presence of such facilitation, agents 1 and 2 can cooperate with each other *as if* they share a friendly social weight τ_1 that allows for an equilibrium with full contributions. Such deal is of course not free and agents

such situation is also possible when social weights attached to other agents are negative ($\tau < 0$). In this case, moral agents will find it socially appropriate to *increase* dissatisfaction of others (e.g., by appropriating or destroying their wealth).

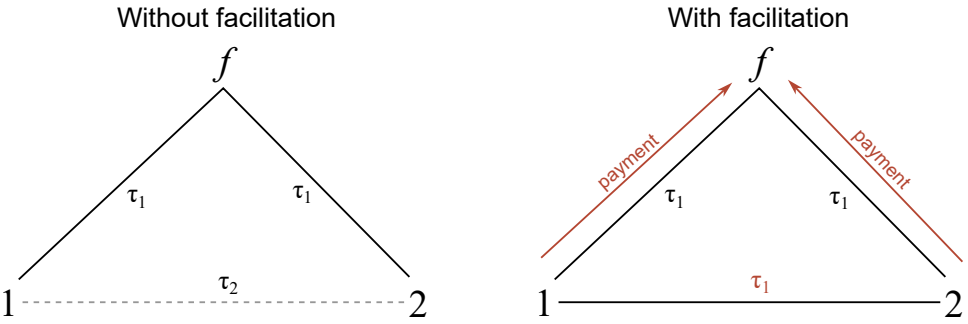

Figure 2: **Left.** Without facilitation players 1 and 2 cannot cooperate due to low social weights τ_2. **Right.** Players 1 and 2 can pay f for facilitation and cooperate with the same social weight τ_1 that they have with the facilitator.

also agree to pay the facilitator f some amount z for his services. Suppose that each agent pays f the amount $z/2$. The right panel of Figure 2 illustrates that now players 1 and 2 can cooperate if they pay the facilitator.

The social dilemma structure of PG ensures that such deals can be profitable for all three agents as long as the payment z to facilitator is "reasonable." In this case, the profit from cooperation for agents 1 and 2 minus the payment $z/2$ can exceed the profit from all other available options, of which there are not so many, and the deal will be made. This kind of interaction can be thought of as the simplest form of *institution for facilitation* that allows for cooperation among strangers and that can emerge spontaneously due to its profitability to all parties involved. Of course, such an institution is not perfect as it cannot connect two agents who do not share a friend. However, with time more complex institutions can evolve from this simple interaction gradually involving more and more unrelated agents.

It is also straightforward to generalize the institution for facilitation to many agents. Suppose that agents in N can interact in some game (Γ, C), but their trust to each other is low so that the Nash equilibrium of (Γ, C) is not on the Pareto frontier (by assumption) and is thus inefficient. By involving a facilitator and paying him z/N each, agents can work in the institution having all trust weights τ_{ij} high enough for all pairs $i, j \in N$ so that Nash equilibrium moves to the Pareto frontier. The gains made in such equilibrium can be enough to cover the payments to the facilitator, and everyone gains as the result.

4.3 Norms in Institutions for Facilitation

In Section 3.2 we analyzed how normativity and the notion of social appropriateness of outcomes emerges in a simple PG. The behavior of moral agents was consequently guided by these notions and allowed us to talk about morally "right" and "wrong" outcomes, or appropriate and inappropriate behavior (which leads to right or wrong outcomes). Using the same technique from KV, we can also analyze social appropriateness of outcomes in the institution for facilitation that includes an additional player, the facilitator. This analysis will allow us to understand the behavior of moral agents as well as define norms that become the inalienable part of this institutional arrangement. Thus, we can calculate, for example, what would be considered reasonable amount of payment z to facilitator that agents 1 and 2 would find "fair."

This amount is the *norm* associated with the institution for facilitation. This is an important notion that leads to the extended welfare analysis for moral agents (with norm-dependent utility) that includes standard consumption considerations as well as their utility losses or gains due to fairness or unfairness of the outcomes resulting from the institution for facilitation [see also Robinson et al., 2023].

To do this, we consider the set of all allocations to three players (agents 1, 2, and f) that can be achieved in some (yet unspecified) non-cooperative game that represents how exactly agents interact in this institution.[10] As in Section 3.2, suppose that players 1 and 2 have endowments w before the game (facilitator does not have or need an endowment since he is not playing the PG). Suppose as well that if each player i pays the amount $z/2$ to facilitator, then both of them have $w - z/2$ left to play the PG. Thus, we can say that the wealth of player i after the game where z was paid to the facilitator is given by

$$w_i^z = w - z/2 - x_i^z + p(x_i^z + x_{-i}^z) \tag{5}$$

where $x_i^z \in [0, w - z/2]$ for both $i \in \{1, 2\}$ is the amount contributed. We define the set of achievable allocations as $C_F = \{(w_1^z, w_2^z, z) \mid z \in [0, 2w] \text{ and } (x_1^z, x_2^z) \in [0, w - z/2]^2\}$. Here we assume that all payments z are possible: from 0 (facilitator works for free) to full endowments of both players $2w$ (facilitator gets all the money and players 1 and 2 have nothing left). The set of allocations C_F is a pyramid in $\mathbb{R}^3$ shown on the left panel of Figure 3.

The definition of C_F is all we need to specify the norm functions of the three players. For player $i \in \{1, 2\}$, we have

$$\begin{aligned}\eta_i(w_i^z, w_{-i}^z, z) &= -[D_i(w_i^z) + \tau_1 D_{-i}(w_{-i}^z) + \tau_1 D_f(z)] \\ &= -\left[\int_{c \in C_F} \max\{c_i - w_i^z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_{-i} - w_{-i}^z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_f - z, 0\} dc\right].\end{aligned} \tag{6}$$

Here we notate $c = (c_i, c_{-i}, c_f) \in C_F \subseteq \mathbb{R}^3$. In words, the social appropriateness of an outcome $(w_i^z, w_{-i}^z, z) \in C_F$ from the perspective of player i is the weighted sum of personal dissatisfactions of the three players, with social weights τ_1 for the other two players (as discussed in Section 4.2).

The norm function for the facilitator f is

$$\begin{aligned}\eta_f(w_1^z, w_2^z, z) &= -[D_f(z) + \tau_1 D_1(w_1^z) + \tau_1 D_2(w_2^z)] \\ &= -\left[\int_{c \in C_F} \max\{c_f - z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_1 - w_1^z, 0\} dc + \tau_1 \int_{c \in C_F} \max\{c_2 - w_2^z, 0\} dc\right].\end{aligned} \tag{7}$$

This is just the reshuffling of the same personal dissatisfactions as above, only with social weights τ_1 applied to the players 1 and 2.

We illustrate these functions in Figure 3. Specifically, the middle panel shows the unique norm func-

[10] An important feature of the framework of KV is that norm functions or measures of social appropriateness are defined on the sets of allocations that ignore the non-cooperative game structure (who moves when, which actions are available, etc.). It allows us to analyze games in the style of cooperative game theory without specifying ex ante how the game unfolds. This property can be very useful for applications where the exact game structure is often unknown or changes depending on circumstances. See Kimbrough and Vostroknutov [2023] for additional discussion.

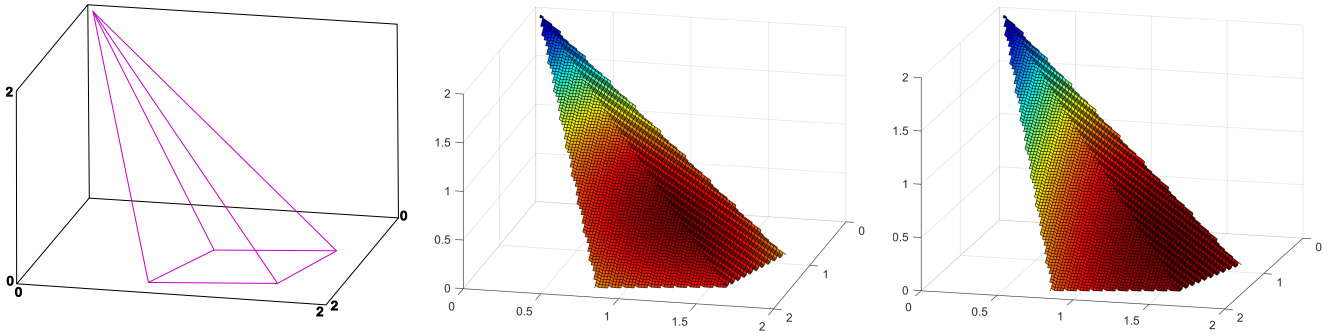

Figure 3: **Left.** The set of allocations C_F (with $w = 1$ and $p = 0.8$). The wealths of players 1 and 2 are on the x- and y-axes; the wealth of the facilitator on the z-axis. **Middle.** The Pareto frontier of the set of allocations C_F in the PG with facilitation. The colors denote the norm function η_i with $\tau_1 = 1$ (dark blue - the lowest social appropriateness; dark red - the highest social appropriateness). **Right.** Same as the middle panel only with $\tau_1 = 0.2$.

tion for all three players that arises when $\tau_1 = 1$, or when both players 1 and 2 trust the facilitator as they trust themselves (e.g., family). This case represents the most cooperative situation achievable within the institution where all players treat each other as they treat themselves (of course, players 1 and 2 do not trust each other much outside the institution for facilitation, they have weight τ_2, see Section 4.2). Notice that in this case the norm (or the most appropriate allocation) obtained from maximization of (7) and marked on the graph in dark red gives 26% of $2w$ to the facilitator ($z = 0.26 \cdot 2w$) and players 1 and 2 should choose to contribute the full amount of what is left after this payment. The right panel of Figure 3 shows the norm function of player 1 for the case when $\tau_1 = 0.2$. In this case, unsurprisingly, player 1 considers it socially appropriate to have more than both other players, which can be seen from the dark red spot being shifted to one side of the pyramid.

The important implication of these calculations, especially that the payment to the facilitator in the most cooperative case is 26% of the players' endowments, is that moral agents can use this as a benchmark to judge how *fair* the payment to the facilitator is. For example, when the facilitator is an actual family member of players 1 and 2 (e.g., brother to one and son-in-law to the other) and demands the payment of, say, 30%, then both players will find it inappropriate and will think that the facilitator is breaking the norm. They might not agree to this arrangement on the basis of their moral convictions and might seek facilitation elsewhere. Similar arguments can be made about taxes and facilitation by the government or some informal institution.

For the general case of game Γ with N players, the set of payoff vectors C, and original endowment $w > 0$ for each player, we can define the set of payoff vectors in the augmented game with the facilitator as $C_F = \{(y, z) \mid z \in [0, Nw] \text{ and } y \in C(w - z/N)\}$, where $C(w - z/N)$ stands for the set of payoff vectors to N players C multiplied by the scalar $(w - z/N)$ or the amount of endowment left after payment

to the facilitator. The norm functions in general case are defined by

$$\eta_i(y,z) = -\left[D_i(y_i) + \tau_1 \sum_{k\in N\setminus i} D_k(y_k) + \tau_1 D_f(z)\right]$$

$$= -\left[\int_{c\in C_F} \max\{c_i - y_i, 0\}dc + \tau_1 \sum_{k\in N\setminus i}\int_{c\in C_F} \max\{c_k - y_k, 0\}dc + \tau_1 \int_{c\in C_F} \max\{c_f - z, 0\}dc\right] \quad (8)$$

for players in N and by

$$\eta_f(y,z) = -\left[D_f(z) + \tau_1 \sum_{k\in N} D_k(y_k)\right]$$

$$= -\left[\int_{c\in C_F} \max\{c_f - z, 0\}dc + \tau_1 \sum_{k\in N}\int_{c\in C_F} \max\{c_k - y_k, 0\}dc\right] \quad (9)$$

for the facilitator.

5 Inclusive and Extractive Institutions

5.1 Game with Facilitator

Now that we know which norms emerge in institutions for facilitation, we can add the non-cooperative game structure to the institution (the explicit specification of the sequence of moves and available actions of the three players) and analyze the various types of equilibrium behavior that result depending on the parameters. In this section, we will assume that there is only one institution for facilitation available [this is relaxed in Robinson et al., 2023] and that players 1 and 2 are "forced" to participate in it, given that without this institution they cannot make any profit at all (they do not trust each other enough to cooperate themselves).[11]

This analysis will allow us to discern two broad classes of equilibria (for different norm-dependent-utility parameters) that can be likened to what literature calls *inclusive* and *extractive* institutions [Robinson and Acemoglu, 2012]. Note that the definitions used in the model are narrower and more model-specific than the broad concepts of inclusive and extractive institutions used in institutional economics. The key properties of these institutions remain the same however: the pluralistic nature of inclusive institutions, where all can participate (at an affordable cost), that leads to more cooperation and production; and exclusive nature of extractive institutions that allow for a group of agents to extract rents above the "fair" cost, from the rest of the population, which results in low production and growth.

[11] This assumption has more evidence to support it than it may seem. In environments similar to Warre, people do not have many options to participate in economic activity and are often forced into extractive institutions. For example, Melnikov et al. [2020] find that significant differences in incomes, dwelling conditions, and employment in large firms, between people living in gang-controlled neighborhoods of El Salvador and those living 50 meters outside of the gang territory are explained by restricted labor mobility enforced by the gangs.

Given that the state of nature does not provide players 1 and 2 with many outside options, the facilitator can exercise power over them since without him players 1 and 2 cannot make any profits. Thus, we define the non-cooperative game structure as follows. First, the facilitator makes players 1 and 2 a take-it-or-leave-it offer of the payment $z \in [0, 2w]$ that he wants for his services. Then, players 1 and 2 pay the requested amount z and play the PG with endowments $w - z/2$. As before, we assume that players 1 and 2 attach social weight $\tau_1 > \tau^*$ to the facilitator (and each other within the institution), so that they optimally cooperate in each subgame happening after any offer z. Notice that in this game all that matters is that players 1 and 2 *believe* that the social weight that they attach to the facilitator and guaranteed by him for their own relationship is τ_1 (high enough for cooperation to be the Nash equilibrium in the subgames). This is important since, in principle, the facilitator does not have to respect or genuinely care about players 1 and 2 at all. All he needs to do for the institution to work (which leads to cooperation and payment of z to himself) is to *convince* the players that he can be trusted.[12]

The last ingredient of the non-cooperative game structure are the utilities that the players eventually receive. This is straightforward: the norm-dependent utility of player $i \in \{1, 2\}$ is given by

$$U_i(x_i; x_{-i}, z) = w_i^z + \phi_i \eta_i(w_i^z, w_{-i}^z, z), \tag{10}$$

and the norm-dependent utility of the facilitator is

$$U_f(z; x_1, x_2) = z + \phi_f \eta_f(w_1^z, w_2^z, z). \tag{11}$$

The analysis of the equilibrium behavior proceeds by backward induction. Given that players 1 and 2 attach high social weight τ_1 to each other within the institution and assuming that their propensities to follow norms ϕ_1 and ϕ_2 are high enough, they will choose full contributions $x_i^*(z) = w - z/2$ given any offer z.[13]

In the next step of backward induction, the facilitator takes the equilibrium actions $x_i^*(z)$ as given and solves the following maximization problem:

$$\max_{z \in [0, 2w]} z + \phi_f \eta_f(p(2w - z), p(2w - z), z).$$

Notice that here we use the wealth of player i given equilibrium play $x_i^*(z)$ and $x_{-i}^*(z)$ determined as $w_i^z = w - z/2 - x_i^*(z) + p(x_i^*(z) + x_{-i}^*(z)) = p(2w - z)$ for $i \in \{1, 2\}$.

What will determine optimal z^* that solves this maximization problem? As follows from the proof of Compromise Theorem [see Proposition 5 in Kimbrough and Vostroknutov, 2022], in the vicinity of the optimum the maximand function is a downward sloping parabola such that $\lim_{\phi_f \to 0} z^* = 2w$ and

[12] In Robinson et al. [2023] we show how this can lead to the emergence of clientelistic networks.

[13] The case when ϕ_1 and/or ϕ_2 are low is uninteresting. Here players 1 and 2 will not contribute anything to the public good and cooperation will fail. Given that in our framework everyone strives to cooperate to make money, we assume that players 1 and 2 are norm-followers to a sufficient degree.

$\lim_{\phi_f \to \infty} z^* = \eta_f^*(w, p, \tau_1)$, where $\eta_f^*(w, p, \tau_1)$ is the maximum of η_f given parameters w, p and τ_1.[14] This means that selfish facilitator with $\phi_f = 0$ will demand the highest possible payment of $2w$ thus leaving players 1 and 2 with nothing, and the most rule-following facilitator with $\phi_f \to \infty$ will ask for the payment that is the most socially appropriate from his perspective (the maximum of η_f) given the optimal actions $x_i^*(z)$ of players 1 and 2 in the subgames. This completely describes the Subgame-Perfect Nash equilibrium in this game. We formulate this result as a proposition.

Proposition 2 (PG). *When ϕ_i are high enough in the PG game with facilitator in SPNE, players 1 and 2 choose to contribute $x_i^*(z) = w - z/2$ or everything they have left after paying any amount z to the facilitator. The facilitator chooses to charge z^* that depends on the parameters, but is ranging from $2w$ (full extraction of resources) to $\eta_f^*(w, p, \tau_1)$, the amount of payment that the facilitator finds mostly socially appropriate or fair.*

Idea of the Proof. We provide full argument in Appendix B. The main difficulty with the proof is to show that the same threshold $\tau^* = (1-p)/p$ applies to Nash equilibria in the subgames of the PG game with facilitator as in PG game without one (see Proposition 1). This needs to be established because the norm function is now computed over the set C_F, and not over individual sets C. But, it turns out that the same proof can be used for this case (given specific structure that within any subgame the payment to the facilitator does not change).

We cannot provide such specific proof for the general case of (Γ, C) where things depend on the properties of its Best-Response correspondence. However following the spirit of the general version of Proposition 1, we can give some characterization of the asymptotic case. Suppose that we have $\phi_i \to \infty$ for all players in N and for the facilitator. Then, their utilities can be described by the general case $\eta_i(y, z)$ and $\eta_f(y, z)$ presented in equations (8-9). In the case with most trust when $\tau_1 = 1$, we have $\eta_i(y, z) = \eta_f(y, z) = \eta(y, z)$, or all the norm functions of all players including the facilitator become identical and so do their utility functions (including the facilitator). In this case on equilibrium path, all players will choose the actions that lead to the most socially appropriate outcome. We formulate this as a proposition.

Proposition 2 (General). *When $\phi_i, \phi_f \to \infty$ and $\tau_1 = 1$ for all $i \in N$, all players including the facilitator will choose actions that lead to the most socially appropriate outcome $(y^*, z^*) = \arg\max_{(y,z) \in C_F} \eta(y, z)$. This will be the unique (cooperative) SPNE of playing (Γ, C) with facilitator. When $\phi_f = 0$, facilitator will extract everything from the players ($z^* = Nw$) in the unique SPNE.*

Idea of the Proof. Same as general Proposition 1. In all subgames, players in N will also choose the unique (by assumption) outcome with the highest social appropriateness given fixed payment z to the facilitator.[15]

[14] By the symmetric nature of η_f (with respect to players 1 and 2), its global maximum coincides with the maximum of $\eta_f(p(2w - z), p(2w - z), z)$ as a function of z.

[15] This proposition suggests that when all players and facilitator trust each other a lot within the institution (high τ) and are all norm-followers (high ϕ), then they can all cooperate and work productively paying some reasonable amounts to the facilitator that everyone finds fair. Moreover the nature of the argument (identical preferences of all extremely norm-following

5.2 Definitions

From the analysis above, it is clear that the "kind" of institution we get in equilibrium depends solely on the payment z^* demanded by the facilitator (since, in general, players fully cooperate in all subgames). Thus, we can classify institutions depending on the value of z^* that also determines the resulting welfare of players in N. When the facilitator is selfish ($\phi_f = 0$), he extracts everything from the players ($z^* = Nw$), which leads to the wealth allocation $(0, ..., 0, Nw)$. This is the worst allocation from the perspective of the welfare of players in N not only because they are left with nothing, but also because this allocation is considered highly inappropriate by all of them. Players in N get the lowest possible level of normative part of their norm-dependent utility at $(0, ..., 0, Nw)$, see Figure 3. We formulate this as a definition.

Definition 1. *We call an institution for facilitation* **fully extractive** *when players in N pay their full endowments w to the facilitator, which gives them the lowest possible level of both consumption and normative utility, and the facilitator gets the highest possible payment Nw.*

Fully extractive institutions can emerge for two different reasons that are related to the two dimensions of trust. The first possibility is when the facilitator is completely selfish ($\phi_f = 0$) and thus disregards the norms: he will act as a standard consumption utility maximizer and extract all surplus from players in N. The second possibility, mentioned above, is that the facilitator is not selfish ($\phi_f > 0$) but cares little for players in N (e.g., his real trust weight towards them is very low). However, facilitator manages to convince players in N that he can be trusted ($\tau_1 > \tau^*$). Then, players in N may agree to participate in the institution, only to learn later that they lose all their endowments.[16]

Notice that fully extractive institutions can emerge for arbitrary levels of trust τ_1 among agents in the institution. This is so because the other dimension of trust, namely the propensity to follow norms ϕ_f, is responsible for creating such conditions, which eliminates the dependency of the equilibrium on trust all together. In order to define the other extreme where the propensity to follow norms is high, we need therefore to take into account both dimensions of trust (ϕ's and τ's). We formulate it as follows.

Definition 2. *We call an institution for facilitation* **fully inclusive** *when $\tau_1 = 1$, or when all players in the institution care about each other as they care for themselves, and when the payment to the facilitator is consistent with the norm and is equal to z^* (from general Proposition 2), or the payment that all three players find most socially appropriate.*

These conditions are reached in equilibrium when $\phi_f \to \infty$, or when the facilitator only cares about social appropriateness and does not care about his own consumption (and when the trust in the institution is the highest, $\tau_1 = 1$).[17]

agents), suggests that in such case even the take-it-or-leave-it structure of the game with facilitator does not play any role: as long as there are no "distortions" in the form of information asymmetries, norms nullify the influence of strategic interactions among players. This is, in a sense, good news, as this suggests not only that things can work given any Γ and any structure of communication with facilitator , but also that we might not need to know the details of these interactions in institutions that work properly.

[16] An example of a fully extractive institution is slavery.

[17] For example, family or a well-functioning democracy are inclusive institutions.

In between fully extractive and fully inclusive institutions lies a continuum of possibilities that are characterized by non-extreme values of the two trust parameters: ϕ_f and τ_1. The problem with classifying this continuum into some flavors of "extractiveness" and "inclusiveness" lies in the fact that for $\tau_1 \in (0,1)$ the norm functions $\eta_i(y,z)$ and $\eta_f(y,z)$ will in general be different and have $N+1$ different maxima (all players in N and f individually believe that they deserve more than the others). Thus, we need to pick a moral perspective from which to judge the welfare qualities of the institution. We propose to take the perspective of the players in N since it is their welfare (and not that of the facilitator) that is important for economic policy.

Suppose that according to the norm function $\eta_i(y,z)$ of player i, the most socially appropriate amount to pay to the facilitator is z_i^*. Then we can provide the following definition.

Definition 3. *For arbitrary level of $\tau_1 \in (0,1)$, call an institution for facilitation* **inclusive** *if the payment to the facilitator is lower than the minimum of z_i^* across all players in N (everyone pays facilitator less than they find fair). If the payment to facilitator is higher than the maximum of z_i^*, then call such institution* **extractive** *(everyone in N pays more than they find fair).*

Notice that in this definition we do not specify what parameters of the norm-dependent utility of the facilitator lead in equilibrium to inclusive or extractive institutions. The reason is that, in general, when $\tau_1 \in (0,1)$ the facilitator who maximizes his norm-dependent utility will always ask for strictly more payment than the maximum of z_i^*. This is so simply because the facilitator cares more about his own dissatisfaction in the norm function $\eta_f(y,z)$ than he cares about the dissatisfactions of players in N. Thus, any informal relationship with $\tau_1 \in (0,1)$ will *always* lead to extractive institutions. Only government that, in principle, does not have to follow its own norm-dependent utility (since it is not a human being) can set the payment to itself to be equal to the minimum of z_i^*, thus creating an inclusive institution.

5.3 Welfare Analysis with Moral Agents

The definitions above characterize institutions in the model and also suggest a new mechanism to increase economic welfare of moral agents that is broader than the standard welfare analysis. Typically in the models with selfish agents (who do not care about norms), the efficient allocation has to be determined by the modeler, or benevolent social planner, and the path to that allocation hinges on absolute market efficiency with the price system that directs selfish agents to the normatively desired allocation. Essentially, such models assume that institutions are already inclusive in the sense that there are no frictions or constraints on the path to the *exogenous* welfare goal. When institutions are extractive, however, markets may not function properly and the path to the efficient allocation may not be possible for selfish agents. This can prevent the achievement of the desired allocation chosen by the social planner within the standard framework.

When morality is introduced into the model, it is *endogenous to agents*, and can therefore describe welfare in the context of both inclusive and extractive institutions. Morality attributed to agents allows

to talk about the allocation that is efficient ("fair") from the perspective of the agents, rather than from the perspective of the social planner. If agents value norm-abiding behavior (they are moral) and can "make the choices they wish" [Robinson and Acemoglu, 2012], then they can achieve this efficient allocation themselves simply because getting there becomes a part of their maximization problem. Agents' morality becomes the driving force towards more inclusive institutions and higher welfare. From this perspective, the role of government or a policy-maker changes to the *creation of the environment* where agents can themselves find a path to inclusive institutions. The measure of inclusiveness, that can be estimated using our definitions, then serves as a policy direction to achieve this goal.

In the world where agents decide themselves what is moral and what is not, the process of increasing welfare is then guided by the goal to make institutions as inclusive as possible (and as little extractive as possible). This can be achieved by making the parameters of the institution close to inclusivity. Policy measures can be taken to increase agents' trust in institutions (high τ_1), to clarify payoffs, rules, and norms (common knowledge of the game; no asymmetric information), and to have facilitators who care about the institution and not only about themselves (high ϕ_f). Notice that all these measures can not only make institutions feel more fair to the agents, but also increase production, growth, and welfare due to higher degrees of cooperation. In other words, they can show moral agents the path towards the narrow corridor.

6 Conclusion

In this paper, we combine the theory of institutions by Acemoglu and Robinson [2020] with the theory of injunctive norms by Kimbrough and Vostroknutov [2022]. In a game-theoretic model with norm-following (moral) agents, we conceptualize institutions as inclusive or extractive. We show how institutions that facilitate cooperation emerge, lead to higher cooperation and can instigate the change from extractive to inclusive institutions.[18] This change leads to higher welfare and allocation of resources that is perceived as "fair" by all agents.

This framework can be used to model any interactions among moral agents, in a wide variety of institutional settings. This opens new possibilities to model institutional change; allows to devise more effective policies by taking into account social norms and institutions; and understand the formation and evolution of institutions in a quantitative framework.

One important direction for future research is to make the framework more practical; to test it with real or experimental/survey data; and to see how to calibrate the model's parameters from surveys or specifically designed tasks. In the next step, policy implications can be considered and estimated, and cost-benefit analysis of impact could be performed. We believe that with a developed applied methodology, our framework can become an indispensable tool for studying institutions, conducting economic policy, and suggesting paths to economic prosperity.

[18] In Robinson et al. [2023] we also consider institutional change resulting from agents' choice between institutions, for example formal and informal; general trust networks; and some sketches of models of historical institutions.

References

D. Acemoglu. Politics and economics in weak and strong states. *Journal of monetary Economics*, 52(7): 1199–1226, 2005.

D. Acemoglu and M. O. Jackson. History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies*, 82(2):423–456, 2015.

D. Acemoglu and J. A. Robinson. *The narrow corridor: States, societies, and the fate of liberty*. Penguin, 2020.

D. Acemoglu, S. Johnson, and J. Robinson. Institutions as the fundamental cause of long-run growth. *Handbook of Economics Growth*, 2005.

D. Acemoglu, S. Naidu, P. Restrepo, and J. A. Robinson. Democracy does cause growth. *Journal of political economy*, 127(1):47–100, 2019.

G. A. Akerlof and R. E. Kranton. Economics and identity. *Quarterly Journal of Economics*, 115(3):715–753, 2000.

K. J. Arrow. Gifts and exchanges. *Philosophy & Public Affairs*, pages 343–362, 1972.

N. Bardsley. Dictator game giving: altruism or artefact? *Experimental Economics*, 11:122–133, 2008.

P. C. Bauer and M. Freitag. Measuring trust. In Uslaner, editor, *The Oxford handbook of social and political trust*, volume 15. Oxford University Press, Oxford, 2018.

C. Bicchieri. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press, 2006.

C. Bicchieri. *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press, 2016.

Y. Chen and S. X. Li. Group identity and social preferences. *American Economic Review*, 99(1):431–57, 2009.

R. H. Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

S. L. Engerman and K. L. Sokoloff. Institutional and non-institutional explanations of economic differences. In C. Ménard and M. M. Shirley, editors, *Handbook of new institutional economics*, pages 639–665. Springer, 2005.

E. Fehr and I. Schurtenberger. Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458–468, 7 2018. doi: 10.1038/s41562-018-0385-5. URL https://rdcu.be/2Jjo.

F. Galeotti, M. Montero, and A. Poulsen. Efficiency versus equality in bargaining. *Journal of European Economic Association*, forthcoming, 2018.

O. Hart. An economist's perspective on the theory of the firm. *Colum. L. Rev.*, 89:1757, 1989.

J. Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2015a.

J. Henrich. Culture and social behavior. *Current opinion in behavioral sciences*, 3:84–89, 2015b.

T. Hobbes. *Leviathan*. Menston, Scolar P., 1651.

B. Holmström and J. Roberts. The boundaries of the firm revisited. *Journal of Economic perspectives*, 12 (4):73–94, 1998.

P. Keefer and S. Knack. Social capital, social norms and the new institutional economics. In *Handbook of new institutional economics*, pages 701–725. Springer, 2008.

P. Keefer and C. Scartascini, editors. *Trust, social cohesion, and growth in Latin America and the Caribbean*. IDB Publications, 2022.

J. B. Kessler and S. Leider. Norms and contracting. *Management Science*, 58(1):62–77, 2012.

E. Kimbrough and A. Vostroknutov. A meta-theory of moral rules. mimeo, Chapman University and Maastricht University, 2023.

E. O. Kimbrough and A. Vostroknutov. The social and ecological determinants of common pool resource sustainability. *Journal of Environmental Economics and Management*, 72:38–53, 2015.

E. O. Kimbrough and A. Vostroknutov. A theory of injunctive norms. SSRN Working Paper, Chapman University and Maastricht University, 2022.

S. Knack and P. Keefer. Institutions and economic performance: Cross-country tests using alternative-institutional measures. *Economics and Politics*, 7(3):207–227, 1995.

E. L. Krupka and R. A. Weber. Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of European Economic Association*, 11(3):495–524, 2013.

K. N. Laland. *Darwin's unfinished symphony: how culture made the human mind*. Princeton University Press, 2018.

S. Lall, U. Akcigit, R. Fattal Jaef, M. M. Ferreyra, K. Karakülah, T. Kleineberg, M. Lebrand, M. Martinez Licetti, D. Merotto, F. Shilpi, K. Stapleton, M. Vagliasindi, and E. Vostroknutova. World development report 2024: The middle-income trap. 2024.

D. Lederman, N. Loayza, and A. M. Menéndez. Violent crime: does social capital matter? *Economic Development and Cultural Change*, 50(3):509–539, 2002.

P. T. Leeson. Better off stateless: Somalia before and after government collapse. *Journal of comparative economics*, 35(4):689–710, 2007.

J. A. List. On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493, 2007.

L.-F. Lopez-Calva, Y. Zhou, E. Al-Dahdah, D. J. Bulman, D. H. Isser, M. Larizza, E. Molina, A. Safir, and S. Sharma. World development report 2017: Governance and the law. 2017.

N. Melnikov, C. Schmidt-Padilla, and M. M. Sviatschi. Gangs, labor mobility and development. Technical report, National Bureau of Economic Research, 2020.

N. Merguei, M. Strobel, and A. Vostroknutov. Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior and Organization*, 197:624–642, 2022.

V. Nee. Norms and networks in economic and organizational performance. *The American Economic Review*, 88(2):85–89, 1998.

D. C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990.

D. C. North. Institutions and the performance of economies over time. In C. Ménard and M. M. Shirley, editors, *Handbook of new institutional economics*, pages 21–30. Springer, 2008.

D. C. North. Institutions and economic theory. *The american economist*, 61(1):72–76, 2016.

E. Ostrom. *Governing the Commons: the Evolution of Institutions for Collective Action*. Political economy of institutions and decisions. Cambridge University Press, Cambridge, 1990. ISBN 9780521405997.

E. Ostrom. Collective action and the evolution of social norms. *Journal of economic perspectives*, 14(3): 137–158, 2000.

F. Panizza, A. Vostroknutov, and G. Coricelli. The role of meta-context in moral decisions. mimeo, Maastricht University, University of Trento, and University of Southern California, 2021.

R. D. Putnam, R. Leonardi, and R. Y. Nanetti. *Making democracy work: Civic traditions in modern Italy*. Princeton University Press, 1992.

J. A. Robinson and D. Acemoglu. *Why nations fail: The origins of power, prosperity and poverty*. Profile London, 2012.

J. A. Robinson, A. Vostroknutov, and E. Vostroknutova. Endogenous institutions and economic policy. The World Bank Policy Research Working Paper no. WPS10600, 2023.

D. Rodrik, A. Subramanian, and F. Trebbi. Institutions rule: the primacy of institutions over integration and geography in economic development. IMF Working Paper No. 02/189, 2002.

M. M. Shirley. Institutions and development. In C. Ménard and M. M. Shirley, editors, *Handbook of new institutional economics*, pages 611–638. Springer, 2005.

A. Vostroknutov. Social norms in experimental economics: Towards a unified theory of normative decision making. *Analyse & Kritik*, 42(1):3–39, 2020.

H. P. Young. Social norms and economic welfare. *European Economic Review*, 42(3-5):821–830, 1998.

Robinson, Vostroknutov, Vostroknutova
"A Model of Endogenous Institutions"

Appendix

A Best Responses in the Public Goods Game

In this appendix we analyze the Nash Equilibria (NE) of the Public Goods game considered in Section 3.2. We assume that the propensities to follow norms ϕ_1 and ϕ_2 are high enough so that the consumption part of the norm-dependent utility (2) is negligible. Thus, we analyze the Public Goods game where the payoffs are equal to the norm function $\eta_i(w_i, w_j; \tau) = -D_i(w_i) - \tau D_j(w_j)$ for player i (we will call j the player who is not i, or $j = -i$).

Our main goal is to prove that there is a threshold value τ^* of trust to the other player such that the Public Goods game with utilities defined by $\eta_i(w_i, w_{-i}; \tau)$ has either 1) unique NE where both players contribute fully (for all $\tau > \tau^*$); 2) unique NE where both players contribute nothing (for all $\tau < \tau^*$); and 3) all pairs of choices of equal contributions are NE (when $\tau = \tau^*$).

We prove this by construction. First, notice that from the proof of the Compromise Theorem [Proposition 5 in Kimbrough and Vostroknutov, 2022] we know that on convex polytopes of allocations—to which class the set of allocations C in the Public Goods game belongs—we can write

$$D_i(w_i) = \sum_{k \in \Omega_{w_i}} a_k^i (b_k^i - w_i)^2 + c_k^i,$$

where Ω_{w_i} is the set of vertices of the polytope that have player i's consumption utility higher than w_i and $a_k^i, b_k^i, c_k^i \in \mathbb{R}$ are some coefficients (they cover all possible quadratic equations). Notice as well that $D_i(w_i)$ is a piece-wise parabola with fixed coefficients for three ranges of w_i. When w_i is high there is only one vertex with higher wealth than that (one set of coefficients). When w_i is lower, there are two vertices with wealth of player i higher than w_i, so the coefficients change. For even lower w_i there are three vertices with higher wealth, so the coefficients change yet again. What is important for us though is the fact that for two players i and j the personal dissatisfaction functions are the same, or that $D_i(w) = D_j(w)$. This comes from the symmetry of the game. Also, since D_i are quadratic, so is η_i, which is a collection of piece-wise-stitched concave quadratic forms.

Now, we want to focus on the allocations that are obtained when both players i and j choose the same contributions $x = x_i = x_j$ that result in some symmetric allocation (w, w). In this case, using Lemma 1 in Appendix C, we have

$$D_i(w) = D_j(w) = a(b - w)^2 + c,$$

where $a, b, c \in \mathbb{R}$ are some coefficients common for both players. These coefficients are also common for all allocations (w, w) because all such allocations have two vertices with wealth larger or same as w (for either player).

Now that this fact is established, we can use it to show that there is a specific value of the trust weight $\tau = \tau^*$ such that the best response of both players with this τ^* is to choose the same contribution as the other player, *no matter what that contribution is*. We show that such τ^* indeed exists and that it also satisfies the first order condition: the derivative at (w, w) along the choices of one player having the other player's action fixed is zero (this guarantees that it is a best response).

Let us write down the derivative. Notice that by definition $\eta_i(w_i, w_j; \tau) = -D_i(w_i) - \tau D_j(w_j)$, which can be rewritten in terms of contribution choices as

$$\eta_i(x_i, x_j) = -D_i(w - x_i + p(x_i + x_j)) - \tau D_j(w - x_j + p(x_i + x_j)).$$

Suppose that x_j is considered fixed and we look at the maximization problem of player i, who chooses x_i to maximize the above (best response). Using the fact that $D_i(w) = D_j(w) = a(b - w)^2 + c$, we can write the first

order condition (the derivative of the above with respect to x_i equal to zero) as

$$-2a(1-p)(b-(w-x_i+p(x_i+x_j)))+2ap\tau(b-(w-x_j+p(x_i+x_j)))=0.$$

Notice that when $x_i = x_j$, or when the players choose equal allocations we are interested in, the big parentheses cancel out and we get the condition

$$-(1-p)+p\tau=0 \tag{12}$$

or

$$\tau=\frac{1-p}{p}=\tau^*.$$

This means that the first order condition above is satisfied for all $x_i = x_j$ only when $\tau = \tau^* = (1-p)/p$. The left panel on Figure 4 illustrates. Here we show the representation of the set of allocations in the Public Goods game (the set of points within the polytope $ABCD$) together with allocations that give both players equal wealth (all in magenta lines). Suppose that x_j is fixed and that player i chooses x_i. Then her choices fall along the lines parallel to AB and DC depending on the value of x_j (AB when $x_j = w$ and DC when $x_j = 0$). The result above means that when $\tau = \tau^*$ the derivatives of η_i are zero for all points (w, w) on the diagonal. This is shown graphically by the dashed black lines (along the edges AB and DC) and little grey dashed lines in between.

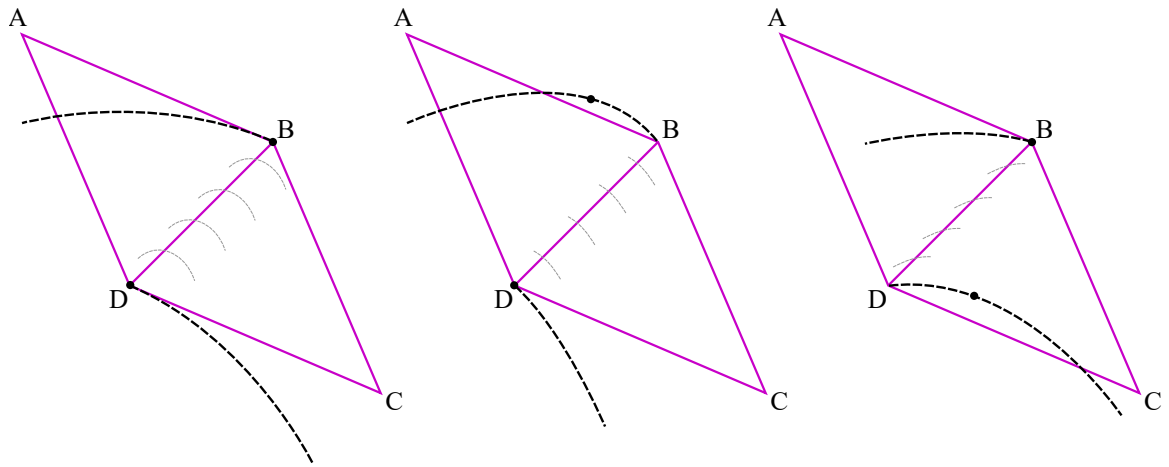

Figure 4: Illustration of best responses in the Public Goods game.

This is also true for the other player. Thus, all this together implies that when $\tau = \tau^*$ the derivatives on the diagonal are zero and are best responses of both players (symmetry). Given that these are mutual best responses, we have a continuum of NE on the diagonal when $\tau = \tau^*$.

This argument demonstrates that there is $\tau^* = (1-p)/p$ such that when the two players with trust τ^* to each other play the Public Goods game, they have a continuum of NE on the diagonal (when they choose the same contributions). This is also a unique τ^* because only it can satisfy the first order condition (12) on the diagonal. The next thing we need to show is that there is a unique NE for all other values of τ, below and above τ^*.

To do that, notice that the condition (12), the left-hand side of which defines the derivative at all diagonal allocations (with equal wealth), tells us that these derivatives are always the same for all such allocations: they are all either positive, negative, or zero depending on the value of τ. For example, the middle panel of Figure 4 shows the case when the derivatives are negative. This means the following. Given that η_i is concave, the negative derivative at point D on the figure implies that it is the maximum on the edge DC (the range of choices of player i given fixed $x_j = 0$) and thus the best response (marked by a black circle). To the contrary, negative derivative at point B means that B cannot be the best response on the edge AB, because the function then grows in the direction of point A. So, the best response is somewhere on the edge AB, but not at B. This last observation also holds in the same way for all points on the diagonal between B and D. Thus, except for the point D, all best responses of player i lie to the left of the diagonal. Similarly, we can work out that for positive derivatives (the

right panel of Figure 4), we have best response at point B and not in D (all best responses in this case are on the right side of the diagonal).

We can use these findings when looking for mutual best responses for the two players. Indeed, when $\tau < \tau^*$ we have the case of negative derivatives. This means that the best responses of the two players are symmetrically situated on the opposite sides of the diagonal except for the point D, where the best responses coincide. They do not coincide anywhere else, since they are divided by the diagonal in all other places. Thus, we can conclude that when $\tau < \tau^*$, we have a unique NE of the game with zero contributions (point D, $x_i = x_j = 0$).

Similarly, when $\tau > \tau^*$ the best responses of the two players coincide at point B, but not anywhere else, since for all other actions the best responses again lie on the opposite sides of the diagonal. Thus, for $\tau > \tau^*$ the unique NE is to contribute fully (point B, $x_i = x_j = w$).

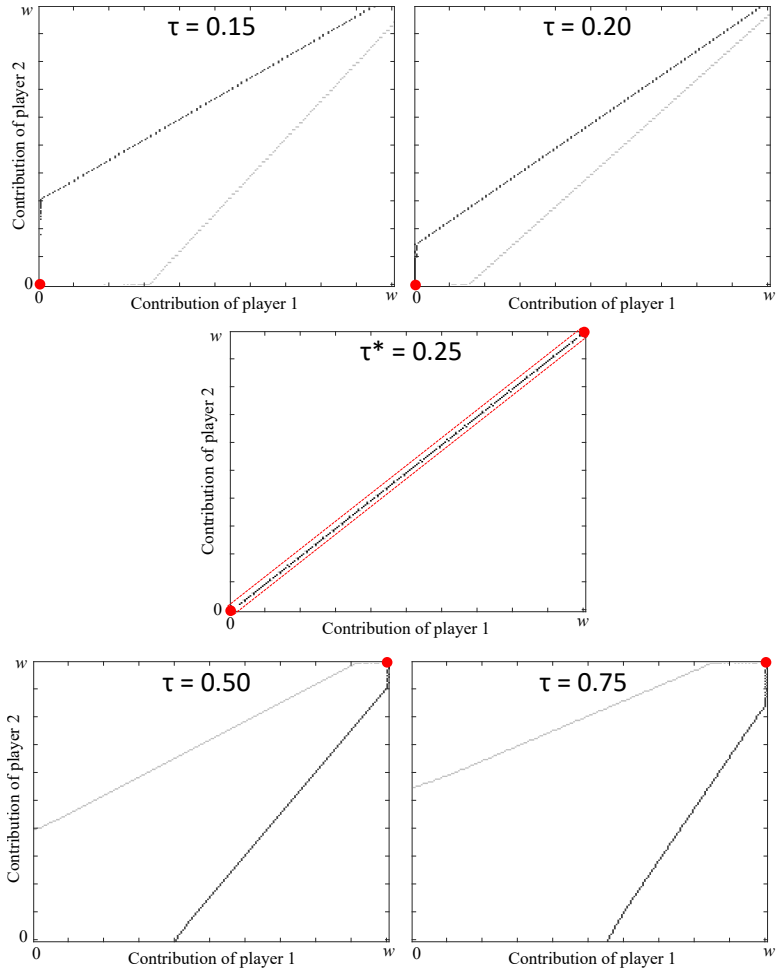

Figure 5: Best response correspondences in Public Goods game with $p = 0.8$ and different levels of τ.

Figure 5 shows numerically computed best response correspondences and Nash equilibria in the Public Goods with $p = 0.8$ for different levels of trust τ of the players to each other. The black lines correspond to the best response of player 1 and grey lines to those of player 2. Equilibria are marked with red circles and dashed lines denote the continuum of equilibria. One can see that for low τ, the unique NE is to contribute nothing as is demonstrated on the top two graphs of Figure 5. For high τ, the NE is to contribute full amounts (the bottom graphs). Finally, when $\tau = \tau^* = 0.25$ we have an intermediate case where any choice of equal contributions constitutes a NE (the middle graph). This same structure of best responses and NE is present for any value of $p \in [0.5, 1)$ when $\tau^* = (1 - p)/p$.

To illustrate, we plot the values of τ^* as dependent on p in Figure 6. Notice that the values on the graph were obtained from the numerical computations of best responses and the resulting values of τ^* are in perfect alignment

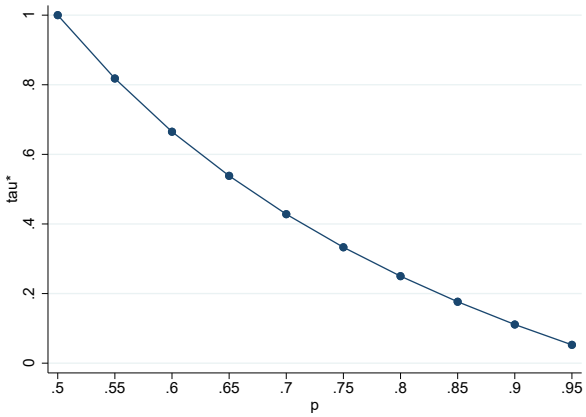

Figure 6: The relationship between productivity p and the cooperation threshold τ^*.

with the function $(1-p)/p$, which validates our computations.

B Best Responses in the Public Goods Game with Facilitation

In this appendix, we show that the threshold property of NE holds also in the Public Goods with facilitation (three players). The logic of this is exactly the same as in the previous section where we considered a simple Public Goods game with two players. To see this notice first that after the facilitator chooses the payment z, the two players find themselves in a subgame, which is essentially a Public Goods game with endowments $w - z/2$ instead of w (z is a given constant). However, we cannot treat this subgame as a simple instance of the Public Goods because the norm function is computed differently using all the allocations in the set C_F (the 3D pyramid on Figure 3).

Despite the complication with the norm function, the logic of the proof of the threshold property stays the same. Indeed, by the results in the proof of Compromise Theorem in Kimbrough and Vostroknutov [2022], we know that for any game with any number of players as long as the allocations are represented by a convex polytope, we have personal dissatisfactions D_i represented by piece-wise connected quadratic functions as in Appendix A. Moreover, by symmetry these functions are the same for players i and j on the diagonal allocations because again they are defined only by the vertices with higher consumption utility and on the diagonal such vertices are always the same for both players (they may differ for different points (w, w), but are the same for a given point (w, w)). This observation—together with the fact that the dissatisfaction of the facilitator in any subgame is constant and can be ignored—allows us to do the same reasoning as in Appendix A and conclude that there is a unique threshold $\tau^* = (1-p)/p$ such that for any $\tau > \tau^*$ the unique NE in all subgames is to contribute fully and for $\tau < \tau^*$ the unique NE in all subgames is to contribute nothing. This result suggests the optimal behavior of the facilitator as described in the main text.

C Lemmata

Lemma 1. *Any non-constant function $f(x) = \sum_k a_k(b_k - x)^2 + c_k$ with some coefficients $a_k, b_k, c_k \in \mathbb{R}$ can be represented as $f(x) = a(b - x)^2 + c$ where $a, b, c \in \mathbb{R}$ are also some coefficients.*

Proof. When we open up the squared terms in f, we get

$$f(x) = (a_1 + ... + a_k)x^2 - 2(a_1b_1 + ... + a_kb_k)x + d,$$

where d is some constant. Then

$$f(x) = (a_1 + ... + a_k)\left[x^2 - 2\frac{a_1b_1 + ... + a_kb_k}{a_1 + ... + a_k}x + \left(\frac{a_1b_1 + ... + a_kb_k}{a_1 + ... + a_k}\right)^2\right] + e,$$

where e is some constant. This can be rewritten as

$$f(x) = a\left(\frac{a_1b_1 + ... + a_kb_k}{a_1 + ... + a_k} - x\right)^2 + c,$$

where $a = a_1 + ... + a_k$ and c is some constant. From the above we have $b = \frac{a_1b_1+...+a_kb_k}{a_1+...+a_k}$. □