The diversity of moral preferences: Evolutionary foundations and some implications in environmental economics

Thèse N° 9478

Présentée le 23 septembre 2019

à la Faculté de l'environnement naturel, architectural et construit
Laboratoire d'économie urbaine et de l'environnement
Programme doctoral en génie civil et environnement

pour l'obtention du grade de Docteur ès Sciences

par

Boris THURM

Acceptée sur proposition du jury

Prof. M. Bierlaire, président du jury
Prof. Ph. Thalmann, directeur de thèse
Prof. I. Alger, rapporteuse
Prof. N. Netzer, rapporteur
Prof. D. Foray, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements - Remerciements

Writing this dissertation has been quite a journey, at times challenging, all the time enriching. I have come a long way, learning a lot and meeting wonderful people. Thus, I am grateful to my supervisor Prof. Philippe Thalmann who offered me this unique opportunity, gave me the freedom to explore topics I am interested in, and trusted me enough to let me work following my (special) rhythm of life.

I would also like to acknowledge the members of my thesis jury, Prof. Ingela Alger, Prof. Nick Netzer, Prof. Dominique Foray and Prof. Michel Bierlaire, who kindly read and evaluated my work, providing insightful feedback and encouraging words for my future career.

Even if pursuing a PhD can sometimes be perceived as a solitary exercise, I was very fortunate to delve into the research universe with Charles Ayoubi. Charles' inputs for this thesis are invaluable. In fact, it is safe to say that this dissertation would have never materialized without him. From Gerzensee classroom to Boston ice rink, I have met more than a talented research companion, I have found a true friend. Merci.

Speaking of Gerzensee, I had the privilege to participate in the 2016 doctoral program. Each trip to the "castle" and its marvelous surroundings was a highlight of my PhD studies thanks to the kindness of the academic and hotel staff, who successfully provided us an ideal learning environment. I enjoyed Gerzensee so much that I never hesitated to return to attend some advanced courses. It is in the Study Center that I achieved the main goal of my doctoral studies, namely to discover economics and to deepen my understanding of this discipline. Thus, I am grateful to the professors who shared with us their knowledge and their passion. In particular, I would like to express my gratitude to Prof. Jörgen Weibull, who inspired most of this thesis and took the time to provide feedback and guidance on my work.

If Gerzensee holds a special place in my heart, it is also and mostly because of the people that I met there. I was lucky to be part of an amazing cohort, so that each week was a treat, full of thrilling moments such as (non-exhaustive list) the "apéros" with Jean-Marc, Andreas' "Gerzensee Fasnacht", Beau's "Zoom Schwartz Profigliano", the minmaxing of Aimilia, Paola's funny stories, the tennis, football, table-football and pool games and the river boat trip. Ideas often emerge from random discussions, and Gerzensee was for sure a place of lively and stimulating discussions. Thanks all for these delightful moments.

Back to Lausanne, I blossomed in the Laboratory of Environmental and Urban Economics. It was my good fortune to share office with Taka. We academically grew up together, always helping each other. Taka, thanks for your kindness, our supportive discussions, and of course

your LaTeX advice. I would also like to acknowledge the comedy duo, Laurence and Marc. Laurence, thanks for all the support and our ski-touring sessions in Verbier. Marc, thanks for your authenticity, for guiding my first steps in academia and for sharing your experience. Likewise, thanks Frank for your advice and wise words. Thanks Mike for our numerous stimulating discussions. Your view on research and your multidisciplinary approach has always been an inspiration. Thanks Margarita for our enjoyable talks and for your delicious cakes, and thanks Sergey for bringing your good mood and leaving us at least a small piece of Margarita's cakes. I also have a thought for Sophie, Vincent, Lucas, Gauthier, Jean-Baptiste and all the others who were part of the LEUrE team at one point or another. Finally, thanks to my new officemate Alex. Replacing Taka was not an easy task, but you did it admirably well, although your Japanese still needs some improvement.

During the last few years, I had the opportunity to be part of the EUCalc project. I have learned a great deal working alongside friendly colleagues, and I would like to acknowledge the EUCalc team for the many interesting and lively discussions. In particular, thank you François and Gino for your humanity, your dedication and for always trying to make the world a better place. François, I also hope that one day I will have your skills to find the most unusual and pleasant "troquets".

Closing this chapter is also the occasion to reflect on my path. From the prépa in Lyon to the SIE Bachelor and the MES Master in Lausanne, my life was shaped by the people I encountered. The list would be too long to name everybody, but I would like to especially thank my accomplices Maad, Kevin and Mehdi for all the good times in Lausanne, Istanbul or Morocco. I also treasure the time I spent in Vancouver during my first academic experience, which convinced me to pursue a PhD.

Si j'en suis là aujourd'hui, c'est aussi, et surtout, grâce au soutien de mes parents. Vous m'avez enseigné le respect et la persévérance, deux qualités essentielles pour la réalisation de cette dissertation. Être têtu m'a grandement aidé à ne pas baisser les bras et à atteindre mes objectifs. Nul doute que vos valeurs d'entraide ont également guidé le choix de mon sujet. Merci du fond du coeur. J'ai aussi une pensée pour ma famille et notamment mes frangins Mathieu, Jean-Ju, Lionel et Pascal qui m'apportent beaucoup de bonheur, comme récemment avec la naissance de mon filleul Raphaël.

Je tiens également à remercier mes amis, qui ont contribué plus qu'ils ne le croient aux succès de cette thèse. En particulier, merci à Dikeus, Polo, Pioup, Tibo, Sacha, Bat, Condo et Vic. Vous rencontrer fût une sacrée bonne idée. Du Yucatán jusqu'en Amérique Centrale, en passant par Lyon, Londres, Porto, Barcelone, et bien sûr Saint-Gervais, ensemble tout est plus joli. J'ai hâte de m'embarquer dans de nouvelles aventures en votre compagnie. Merci aussi à Kekette pour nos voyages dans l'espace jusqu'au bois des fées. Un très grand merci à Lolo. Ton optimisme, ton côté bon vivant, ta bonne humeur en toute circonstance (ou presque), ta confiance en toi et ta persévérance ont toujours été une grande source d'inspiration et de motivation pour moi. Merci aux autres boutch de Cham, notamment Antoine, Coco, Krichka, Papaye, Rémy, Ben, Vincent et le seul et unique Valou. Des galops chez Gilles à la gnôle de fin de soirée, votre gentillesse, votre folie, votre simplicité et votre modestie m'ont sans cesse permis de

faire la part des choses et de garder le sens des priorités. Merci à Sylvie et Didier, mes saisons à la Crêmerie ont été une formidable école de la vie. Je suis également très reconnaissant envers Damien, dis "Dàdùzi" dans certaines contrées, qui a en quelque sorte été mon mentor. Tes conseils et ton expérience ont été d'une aide précieuse, ton goût de la précision m'a constamment poussé à améliorer la qualité de mon travail, et nos randos et sessions de grimpe m'ont offert des pauses revigorantes. Pour finir, comment ne pas mentionner celui qui m'accompagne et me supporte depuis nos 6 ans et les bancs de l'école primaire. Jérém, tu m'as toujours tiré vers le haut (alors qu'au hockey le tir n'a jamais vraiment été ta spécialité...), tu m'as soutenu dans les bons et les mauvais moments, et tu es à l'origine de cette dissertation puisque c'est grâce à toi que je suis venu à l'EPFL. Je suis fier de notre amitié.

Last but not least, I owe a debt of gratitude to Sophia for her support, for her proofreading, for taking such a good care of me in sometimes stressful times, and also for helping my dream of seeing elephants and lions come true. Thanks for your kindness, for accepting me just the way I am, and for being who you are. Even if you still have a long way to go to improve my English (let alone my German), you made me a better person and there are no words to express how grateful I am to you.

Lausanne, le 9 Mai 2019 Boris Thurm

Abstract

Empirical evidence suggests that there exists substantial heterogeneity in individuals' social preferences. However, there is little theoretical basis supporting this observation and economic models often assume that all individuals are identical. Hence, the aim of this thesis is to provide theoretical foundations for the observed heterogeneity of social preferences and to derive its implications for environmental policy. We first extend the framework of evolutionary game theory introducing the concepts of *evolutionarily stable population* and of assortment matrix to study the evolution of preferences in assortatively matched interactions between heterogeneous individuals. We show that there exists a heterogeneous *evolutionarily stable population* composed of both fully-selfish and fully-moral individuals for some but not all games and assortment structures. Therefore, the preferences that are favored by evolution depend on the socio-economic environment. In particular, our analysis highlights the key role played by the assortment structure in the existence and the robustness of heterogeneous *evolutionarily stable populations*. We then design a model with heterogeneous moral individuals involved in a social dilemma. Our framework sheds light on many empirical findings explaining why some individuals are willing to voluntarily engage in costly pro-environmental actions even though the impact of their efforts on environmental externalities is negligible. Investigating how individuals' beliefs can alter their behaviors and hinder cooperation, we demonstrate why financial incentives can fail to foster pro-environmental behaviors in some cases while non-financial incentives such as nudges and educational campaigns could be successful. Consequently, better accounting for the social motives behind individuals' decisions in economic models could help policy makers design more effective policies.

KEYWORDS: Social preferences, Heterogeneity, Morality, Homo moralis, Cooperation, Preference evolution, Evolutionary game theory, Assortative matching, Social Dilemma, Environmental policies

Résumé

Les données empiriques suggèrent que les préférences sociales des individus sont très hétérogènes. Cependant, il n'y a que peu d'analyses théoriques expliquant cette observation. De plus, les modèles économiques font souvent l'hypothèse que tous les individus sont identiques. L'objectif de cette thèse est de remédier à ces manques en apportant des fondements théoriques à l'hétérogénéité observée des préférences sociales, et en analysant les implications de cette diversité pour les politiques environnementales. Tout d'abord, nous étudions l'évolution des préférences dans une population dans laquelle les individus possèdent diverses préférences sociales. Pour ce faire, nous élargissons le cadre de la théorie des jeux évolutionnaires en introduisant le concept de *population évolutivement stable*. Dans une telle population, des individus hétérogènes coexistent et résistent à l'invasion d'un petit groupe d'individus qui ont une préférence différente. L'appariement entre les individus est assortatif : deux individus partageant la même préférence ont davantage de chances de se rencontrer que deux individus ayant des préférences distinctes. Nous montrons qu'il existe une *population évolutivement stable* constituée d'individus égoïstes et moraux pour certains jeux et structures d'appariement mais pas pour tous. Ainsi, les préférences favorisées par l'évolution dépendent du contexte et de l'environnement socio-économique. En particulier, notre analyse met en évidence le rôle clé joué par la structure d'appariement dans l'existence et la robustesse d'une *population évolutivement stable*. Nous étudions ensuite le comportement d'individus diversement moraux qui interagissent dans un dilemme social. Notre modèle permet d'expliquer pourquoi certaines personnes sont disposées à effectuer des actions en faveur de l'environnement, même si leurs efforts ont un effet négligeable sur l'externalité environnementale. Nous examinons comment une perception erronée peut modifier les choix des individus et entraver le développement de la coopération dans la population. Nous montrons également pourquoi les incitations financières peuvent échouer à promouvoir des comportements respectueux de l'environnement, tandis que des incitations non financières telles que des campagnes de sensibilisation et des certificats verts peuvent être couronnées de succès. Cette thèse démontre que la prise en compte dans les modèles économiques des motivations non pécuniaires influant les décisions des individus peut aider les décideurs à concevoir des politiques plus efficaces.

MOTS CLÉS : Préférences sociales, Hétérogénéité, Moralité, Homo moralis, Coopération, Evolution des préférences, Théorie des jeux évolutionnaires, Appariement assortatif, Dilemme social, Politiques environnementales

Contents

List of Figures

List of Tables

1 Introduction

Although commonly used in the economics literature, the *homo oeconomicus* hypothesis of rational agents pursuing their self-interest fails to explain many human behaviors (Henrich et al., 2001). For instance, empirical evidence suggests a consistent tendency by some individuals to cooperate in public good games (see e.g. Marwell and Ames, 1981; Fischbacher et al., 2001; Brekke et al., 2011). As shown by Andreoni (1995), this propensity to contribute to the public good is robust and cannot be blamed on the agent's confusion. Hence, ever since Smith (1759) suggested moral motives in his *Theory of moral sentiments*, economists have considered several alternative preferences such as altruism (Becker, 1974b), warm glow (Andreoni, 1990), fairness (Rabin, 1993), empathy (Stark and Falk, 1998), reciprocity (Fehr and Gächter, 1998), reciprocal altruism (Levine, 1998), inequity aversion (Fehr and Schmidt, 1999) or morality in the Kantian sense[1] (Laffont, 1975; Brekke et al., 2003; Alger and Weibull, 2013).

In a recent study, Falk et al. (2018) have analyzed the global variation in social preferences such as altruism, trust, reciprocity, time and risk preferences. Their analysis reveals *"substantial heterogeneity across countries, but even larger within-country heterogeneity"* (Falk et al., 2018, p. 1645). This diversity has been observed in other contexts such as voting behavior (Piketty, 1995), environmental consciousness (Schlegelmilch et al., 1996) and willingness-to-pay for climate change mitigation (Layton and Brown, 2000). While understanding the mechanisms behind this heterogeneity is crucial to better comprehend individuals' decision making, there is a lack of theoretical studies analyzing the origin of this diversity. Furthermore, economic models often assume that all individuals are identical for simplicity. But overlooking the heterogeneity of social preferences could have important policy implications, for example in the presence of externalities (Kaplow, 2008).

Consequently, the aim of this thesis is to provide theoretical foundations for the observed heterogeneity of social preferences and to derive its policy implications in the context of environmental, energy and resource economics.

[1] Kant (1870) first formulation of his categorical imperative is: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.".

In Chapter 2, we examine the theoretical basis of heterogeneous social preferences by extending the framework of evolutionary game theory to account for a diversity of preferences in the population. Inspired by the work of Alger and Weibull (2013, 2016) who proved that moral preferences are favored in homogeneous populations, we show that fully-selfish *homo oeconomicus* individuals and fully-moral *homo kantiensis* individuals can coexist in a population and be favored by evolution in a heterogeneous population. Conversely to the classical setting, we find that the favored preferences in a heterogeneous population are context-dependent.

In Chapter 3, we design a model with heterogeneous moral agents in order to better understand individuals' decision making in social dilemmas. We analyze why some (but not all) individuals are willing to engage in costly pro-environmental actions even though their efforts have a negligible impact on the environmental externality. We also explore the influence of individuals' beliefs on their decisions. We provide several applications to illustrate how our model can shed light on much empirical evidence. Last but not least, we discuss the policy implications of accounting for a heterogeneous moral population. In particular, we show why financial incentives could fail in some cases while relying on other instruments such as nudges and education can be effective.

Finally, we conclude in Chapter 4 summing up our findings and discussing future work.

2 Exploring the diversity of social preferences

Disclaimer: This chapter draws from several working papers written with Charles Ayoubi. I would like to express my gratitude to Charles for his contribution and our enjoyable countless discussions. This chapter would have never materialized without him. I would also like to acknowledge Prof. Jörgen Weibull, Prof. Ingela Alger, Prof. Martin Nowak, Prof. Klaus Schmidt, Prof. Philippe Thalmann, Dr. Fabiana Visentin, Dr. Damien Ackerer and Sophia Ding for their valuable feedback.

2.1 Motivation

The Global Preferences Survey reveals that individuals exhibit considerable differences in their social preferences (Falk et al., 2018). The global variation in levels of altruism is for instance illustrated in Figure 2.1. The findings of Van Leeuwen et al. (2012), showing that chimpanzees also exhibit a diversity of social behaviors, hint at the possibility of an evolutionary origin behind this heterogeneity. Our goal in this chapter is to assess the evolutionary foundation of the coexistence of more than one type of preference in a population, and to evaluate what types of preferences prevail then.

Our analysis is inspired by the work of Alger and Weibull (2013, 2016). In a model of preference evolution under incomplete information and assortative matching, they show that a new type of preference, called *homo moralis*, arises endogenously as the most favored by evolution. A *homo moralis* individual maximizes a weighted sum of her selfish payoff and of her moral payoff, defined as the payoff that she would get if everybody acted like her.[1] The *homo moralis* preferences elegantly tackle the shortcomings of selfish preferences. However, by building on the classical definition of evolutionary stability by Maynard Smith and Price (1973), Alger and Weibull (2013, 2016) investigate the survival of only one type of preference in the population.

When Maynard Smith and Price (1973) and Maynard Smith (1974) laid the foundations of

[1] Bergstrom (1995) also showed the evolutionary stability of a "semi-Kantian" utility function (a *homo moralis* with morality coefficient one half) in the special case of symmetric interactions between siblings.

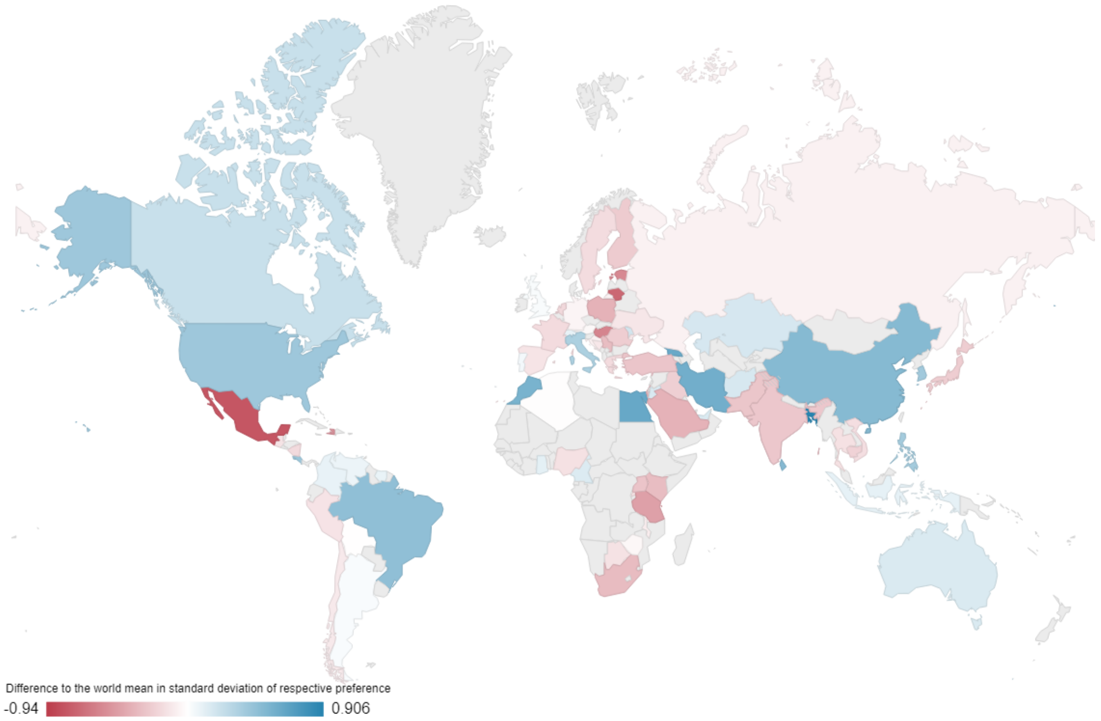

Difference to the world mean in standard deviation of respective preference
-0.94
0.906

Figure 2.1 – Global variation in altruism. Source: Falk et al. (2018). Global evidence on economic preferences. *QJE*, 133(4), 1645-1692. More information on the Global Preference Survey: https://www.briq-institute.org/global-preferences

evolutionary game theory, they aimed at identifying the strategy providing an evolutionary advantage in animal conflicts between members of a given species. Therefore, they defined the concept of an *evolutionarily stable strategy*, a strategy adopted by most of the members of a population (called the "resident" strategy), which give a higher reproductive fitness than any other "mutant" strategy. Alger and Weibull (2013) generalize this definition of evolutionary stability, applying it to preference evolution, in order to identify an *evolutionarily stable preference*. A peculiar *homo moralis* type of preference emerges in this framework as evolutionarily stable under assortative matching. However, assuming that all resident individuals have the same preference, their approach abstracts from the empirically observed heterogeneity of preferences among individuals. Our aim is to fill this gap.

We first introduce the concept of an *evolutionarily stable population* defining the conditions under which several types can coexist in a population and resist a small-scale invasion of any other type. We also design an assortment matrix to portray assortatively matched interactions between individuals. We then analyze the evolutionary stability of a heterogeneous population composed of two types of *homo moralis*, the fully-selfish *homo oeconomicus* and the fully-moral *homo kantiensis*, involved in a social dilemma. We show that there exists a heterogeneous *evolutionarily stable population* for some but but not all games and assortment structures, and we characterize the conditions for this existence.

Our work contributes to the literature on the evolution of preferences. When preferences are unobservable, selfish motives prevail in large groups of uniformly-matched individuals (Ok and Vega-Redondo, 2001; Dekel et al., 2007). On the other hand, two main drivers favoring the evolutionary success of other social preferences have been identified in the literature. First, when opponents' preferences are (at least partly) observable, evolution can lead to the emergence of altruism, reciprocal behaviors or spiteful preferences (Bester and Güth, 1998; Fershtman and Weiss, 1998; Koçkesen et al., 2000; Heifetz et al., 2007; Herold, 2012). Second, and as discussed above, *homo moralis* has an edge when the matching process is assortative (Alger and Weibull, 2013, 2016). However, most of these papers assume a homogeneous resident population. In contrast, we consider a heterogeneous resident population.

Our work also relates to the literature on the evolution of cooperative behaviors. The evolution of strategies under assortative matching has been extensively studied (mostly in biology) in the context of evolutionary game dynamics.[2] For example, Bergstrom (2003), Allen and Nowak (2015) and Jensen and Rigos (2018) explore the evolution of cooperation in social dilemmas.[3] Their findings are in line with ours when the cooperating individuals are represented by the fully-moral *homo kantiensis* preference and the defectors by the fully-selfish *homo oeconomicus* preference. Nonetheless, we go one step further in the analysis by determining circumstances under which the population can resist the invasion of mutants. As we will see later on, not all heterogeneous populations at the equilibrium in a dynamic setting can actually withstand the mutants' invasion.

Finally, since our population consists of two resident types and one mutant type, we generalize the algebra of assortative matching previously derived by Bergstrom (2003, 2013) for encounters between two types. In our model, individuals interact in pairs. Recently, Jensen and Rigos (2018) study the more general case of encounters in groups of various sizes. While they focus on assortment between strategies and define a matching rule (specifying how individuals playing different strategies are allocated into groups of various sizes) to obtain the matching probabilities, we introduce a type-by-type assortment matrix which characterizes the assortment between preferences. Still, the two approaches are closely linked and result in similar outcomes.

The organization of the rest of the Chapter is as follows. In Section 2.2 we present the model and the main definitions, introducing the assortment matrix and the concept of *evolutionarily stable population*. In Section 2.3 we analyze the evolutionary stability of a heterogeneous population composed of *homo oeconomicus* and *homo kantiensis* individuals. In Section 2.4 we review the differences between homogeneous and heterogeneous *evolutionarily stable populations*. In Section 2.5 we allow for a greater diversity, discussing the evolutionary stability of population composed of other types than *homo oeconomicus* and *homo kantiensis*. Finally, we recap our findings and we suggest potential extensions in Section 2.6.

[2] See for instance Hofbauer and Sigmund (2003), Sandholm (2010) and Nowak et al. (2010) for a description and review of the field; and also Nowak (2006) for a discussion of mechanisms allowing the survival of cooperation.

[3] Bilancini et al. (2018) also look at the evolution of cooperation between assortatively matched individuals, introducing heterogeneity in culture, to investigate the effect of cultural intolerance.

2.2 Model and definitions

In this section, we present the model, the assumptions made and the main definitions. We consider a large population of individuals of different types, i.e. preferences (Section 2.2.1). Individuals interact in pairs and the matching is assortative (Section 2.2.2). While individuals' behaviors are driven by their preferences, their evolutionary success is determined by the payoffs they get (Section 2.2.3). We introduce the concept of *evolutionarily stable population* in Section 2.2.4 and a particular type of preference, *homo moralis*, in Section 2.2.5. Finally, throughout most of this chapter, we will analyze the evolutionary stability of a population of two types of *homo moralis*, namely *homo oeconomicus* and *homo kantiensis*, involved in a prisoners' dilemma (Section 2.2.6).

2.2.1 Heterogeneous Population

We consider a large population of individuals whose behaviors depend on their type $\theta_i \in \Theta$, i.e. their preferences. In the classical setting, a population is composed of two types $(\theta_1, \theta_\tau) \in \Theta^2$ (Alger and Weibull, 2013). The two types and their respective shares define a population state $s = (\theta_1, \theta_\tau, \lambda_\tau)$, where $\lambda_\tau \in (0,1)$ is the population share of θ_τ. If λ_τ is small, θ_1 is called the resident type and θ_τ the mutant type.

We expand the classical model by allowing for the presence of three types $(\theta_1, \theta_2, \theta_\tau) \in \Theta^3$. Let $I = \{1, 2, \tau\}$, then for all $i \in I$, we denote the share of type θ_i in the population by $\lambda_i \in (0,1)$. The three types and their respective shares define a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$. By normalizing the population size to unity, we have: $\sum\limits_{i \in I} \lambda_i = 1$. Therefore, the population state s could be described with only two population shares instead of three. For convenience, we will often use $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with λ the relative share of θ_2 with respect to θ_1, i.e. $\lambda = \lambda_2/(\lambda_1 + \lambda_2)$. Note that we have: $\lambda_1 = (1-\lambda)(1-\lambda_\tau)$ and $\lambda_2 = \lambda(1-\lambda_\tau)$.

When λ_τ is small, i.e. when $\lambda_\tau << \lambda_1$ and $\lambda_\tau << \lambda_2$, θ_1 and θ_2 are called the resident types and θ_τ the mutant type.[4] A population with at least two resident types is called *heterogeneous*, while a population with one resident type is called *homogeneous*.

2.2.2 Matching

Individuals are randomly matched into pairs. For all $(i, j) \in I$, the conditional probability that an individual of type θ_j is matched with an individual of type θ_i is called $p_{i|j}$.[5] The matching process is exogenous[6] and it may be assortative.

[4] By extension, we will sometimes talk about residents (mutants) to refer to the individuals of the resident (mutant) type.

[5] Note that all the probabilities are a function of the population state s but we drop this precision for readability purposes.

[6] Allowing individuals to select their partners (Becker, 1973, 1974a; Gunnthorsdottir et al., 2010; Jackson and Watts, 2010) would require to include informational and strategic features beyond the scope of this study.

Assortative Matching

In a situation of assortative matching, the probability to meet an individual of type θ_i is not necessarily the same for an individual θ_i and for an individual θ_j, i.e. we can have $p_{i|i} \neq p_{i|j}$. This contrasts with the case of uniform-random matching in which the probability to meet an individual of type θ_i is always equal to the share λ_i of θ_i in the population, i.e. for all $(i,j) \in I$, $p_{i|j} = p_{i|i} = \lambda_i$.

In the classical setting with two types in the population, Bergstrom (2003) introduced an assortment function in order to model assortative encounters. Building on his approach, we introduce the concept of a type-by-type assortment matrix function allowing for assortative matching in interactions between individuals of three distinct types.

Definition 1 (Assortment matrix)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, for all $(i,j) \in I^2$, let $\phi_{ij}(\lambda, \lambda_\tau)$ be the difference between the conditional probability to be matched with type θ_i, given that the individual herself is of type θ_i, and the probability to be matched with type θ_i, given that the individual is of type θ_j: $\phi_{ij}(\lambda, \lambda_\tau) = p_{i|i} - p_{i|j}$.
For all $(i,j) \in I^2$, $\phi_{ij} : (0,1)^2 \to [-1,1]$. This defines an exogenous assortment functions matrix: $\Phi = ((\phi_{ij}(\lambda, \lambda_\tau)))_{(i,j)\in I^2}$.

Extending the concept of assortment function, the assortment matrix embeds *homophily* effects, i.e. the tendency of individuals to interact more with others with similar characteristics such as family, ethnicity, age, gender, language, religion, geographic proximity, education, work, association activity or income (Ibarra, 1993; McPherson et al., 2001). The assortment matrix allows accounting for the higher probability of interacting with similar others (Byrne, 1971; Lakin and Chartrand, 2003), relating to the notion of distance in network economics (Currarini et al., 2009; Iijima and Kamada, 2017). Some alternative approaches to model *homophily* in an evolutionary framework include evolutionary graph theory and evolutionary set theory (Nowak et al., 2010). In the former, individuals occupy the vertices of a graph and their interactions are governed by edges (Lieberman et al., 2005; Ohtsuki and Nowak, 2008; Shakarian et al., 2012). In the latter, individuals belong to several sets (e.g. school, company, living location, associations, etc.) and the more sets they have in common, the more interactions exist between them (Tarnita et al., 2009). The assortment matrix defined above is exogenous and hence allows for large flexibility in the setting of the assortment as a function of the state s. It can therefore be used in a variety of contexts like economics, sociology, biology or management, with the possibility to calibrate its values empirically.

We now introduce a particular type of assortment matrix extending the classical case of constant assortment often used in single-resident populations (Alger and Weibull, 2012; Salmon and Wilson, 2013) derived from the Wright's coefficient of relatedness in biology (Wright, 1922). This definition will be useful in the evolutionary stability analysis in Section 2.3.

Definition 2 (Uniformly constant assortment matrix)**.** An assortment matrix Φ is called *uniformly constant* when all of its non-diagonal components are independent of the population shares and equal to the same value.[7] In other words, we will say that Φ is *uniformly constant*[8] when, for all $(i, j, k, l) \in I^4$ such that $i \neq j$ and $k \neq l$:

$$\begin{cases} \phi_{ij} : (0,1)^2 \to [-1,1] \quad \text{is} \quad \text{constant}, \\ \phi_{ij}(\cdot) = \phi_{kl}(\cdot) \end{cases}$$

Note that the case of uniform random matching is a special case of uniformly-constant assortment where each assortment function is constant and equal to zero: $\Phi = ((0))_{(i,j)\in I^2}$.

We assume that for all $(i, j) \in I^2$, $\phi_{ij}(\cdot)$ is continuous in its two arguments (λ, λ_τ) and converges as the mutant share in the population λ_τ goes to zero. We define the assortativity σ:

Definition 3 (Assortativity)**.** The assortativity $\sigma \in [0,1]$ is the limit for all $i \in \{1,2\}$ of $\phi_{\tau i}$ when λ_τ goes to zero:

$$\forall\, i \in \{1,2\} : \quad \lim_{\lambda_\tau \to 0} \phi_{\tau i}(\lambda, \lambda_\tau) = \sigma$$

Using the definition of assortativity, the assortment functions $\phi_{ij} : (0,1)^2 \to [-1,1]$ can be extended by continuity to $(0,1) \times [0,1)$ to cover the limit when the mutant share θ_τ goes to zero. We will also note $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ the population state when the mutant share goes to zero.

Remark 1. At the limit when λ_τ goes to zero, we have for all $i \in \{1,2\}$, $\phi_{\tau i}(\lambda, 0) = \sigma = p_{\tau|\tau}$. Indeed, according to the balancing conditions (see Property 2 below), the probability for a resident to be matched with a mutant $p_{\tau|i}$ is zero. Thus, the assortativity is independent of the resident types, and we also have $\sigma \in [0,1]$.

Remark 2. The continuity of the assortment functions and the definition of assortativity $\sigma \in [0,1]$ imply that any uniformly-constant assortment matrix can be written as a function of the unit-matrix J[9] and the identity matrix I as follows: $\Phi = \sigma(J - I)$.

Matching probabilities

The matching process must satisfy some properties in order to be well defined. We detail these properties in this section and show how the matching probabilities can be written only in function of the population shares and the assortment matrix. In the following, we will use the notation ϕ_{ij} to designate $\phi_{ij}(\lambda, \lambda_\tau)$, abstracting from the arguments of the assortment functions for simplicity.

[7] By definition of the assortment functions, the matrix Φ has a diagonal of zeros.

[8] By extension, we will say that the assortment is *uniformly constant* when the assortment matrix is *uniformly constant.*

[9] The unit-matrix J is the matrix having each of its components equal to one.

Property 1 (Matching conditions)**.** The conditional probabilities satisfy the matching conditions if each individual is matched with another individual with probability one, i.e. nobody is left behind without a match:

$$\forall\, i \in I: \quad \sum_{j\in I} p_{j|i} = 1$$

Property 2 (Balancing conditions)**.** The conditional probabilities satisfy the balancing conditions if the probability of the event "being of type θ_i and being matched with an individual of type θ_j" is the same as the probability of the event "being of type θ_j and being matched with an individual of type θ_i":

$$\forall\, (i,j) \in I^2: \quad \lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$$

The balancing conditions ensure the coherence of the matching process. Similarly, in order to be well defined, the assortment matrix must satisfy some conditions that we call the assortment balancing conditions:

Property 3 (Assortment balancing condition)**.** The assortment matrix satisfies the *assortment balancing conditions* when:

$$\forall\, (i,j) \in I^2: \quad \lambda_j \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{ik}\right) - \phi_{ij}\right] = \lambda_i \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{jk}\right) - \phi_{ji}\right]$$

If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions.

Proof. In Appendix A.1. □

The assortment balancing conditions impose a particular relationship between the assortment functions. As noted by Bergstrom (2003) in the case of assortative encounters between two types, the assortment $\phi_{12} = p_{1|1} - p_{1|2}$ defined between a type θ_1 and a type θ_2 is equal to the assortment $\phi_{21} = p_{2|2} - p_{2|1}$ defined between θ_2 and θ_1. When a third type θ_τ is part of the population, this result does not hold anymore, i.e. we do not necessarily have $\phi_{12}(\lambda,\lambda_\tau) = \phi_{21}(\lambda,\lambda_\tau)$. However, at the limit when the mutant share goes to zero, the residents are matched between them, as if there was no mutants, and thus we get the same relation $\phi_{12} = \phi_{21}$. Formally:

Lemma 1 (Assortment between residents)**.** *When $s = (\theta_1,\theta_2,\theta_\tau,\lambda,0)$, if the matching process satisfies the matching and balancing conditions, then $\phi_{12}(\lambda,0) = \phi_{21}(\lambda,0)$.*

Proof. In Appendix A.2. □

Knowing the assortment matrix Φ, we have a system of equations on the conditional probabilities $p_{i|j}$ defined by:

- The matching conditions: for all $i \in I$, $\sum_{j \in I} p_{j|i} = 1$ (Property 1)
- The balancing conditions: for all $(i, j) \in I^2$, $\lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$ (Property 2)
- The assortment matrix conditions: for all $(i, j) \in I^2$, $\phi_{ij} = p_{i|i} - p_{i|j}$ (Definition 1)

When the assortment matrix satisfies the assortment balancing conditions, this system has a unique solution, i.e. we can express the conditional probabilities in function of the population shares and assortment functions:

Proposition 1 (Matching probabilities)**.** *When the assortment matrix Φ satisfies the assortment balancing conditions (Property 3), the system defined by matching conditions (Property 1), balancing conditions (Property 2) and assortment matrix conditions (Definition 1) has a unique solution:*

$$\forall (i, j) \in I^2: \quad p_{i|j} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \tag{2.1}$$

Proof. In Appendix A.3. □

Property 4. Since for all $(i, j) \in I^2$, $p_{i|j} \in [0,1]$, the assortment functions should respect another set of conditions to be coherent with the matching process:

$$\forall (i, j) \in I^2: \quad 0 \leq \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik} - \phi_{ij} \leq 1$$

Remark 3. Note that under uniform random matching, for all $(i, j) \in I^2$ $\phi_{ij} = 0$ and we obtain $p_{i|j} = \lambda_i$, i.e. each individual is matched with an individual of type θ_i according to the population share λ_i of individuals of type θ_i.

It is also interesting to detail the conditional probabilities $p_{i|i}$:

$$\forall\, i \in I: \quad p_{i|i} = \lambda_i + \sum_{k \in I} \lambda_k \phi_{ik}$$

The conditional probabilities $p_{i|i}$ are the sum of several terms. The first, λ_i, is the population share of individuals of type θ_i. The others, $\lambda_k \phi_{ik}$, represent the additional matching between individual of type θ_i at the expense of matching with individuals of type θ_k, weighted by λ_k the population share of individuals of type θ_k.

Finally, we will need to know the limits of the conditional probabilities when the mutant share λ_τ goes to zero.

Lemma 2 (Matching probabilities in a population of two residents and one mutant)**.** *When $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, if Proposition 1 is satisfied, then we have:*

$$\begin{aligned}
p_{1|1} &= (1-\lambda) + \lambda \cdot \phi_{12} \\
p_{1|2} &= (1-\lambda) \cdot (1-\phi_{12}) \\
p_{1|\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{2|1} &= \lambda \cdot (1-\phi_{12}) \\
p_{2|2} &= \lambda + (1-\lambda) \cdot \phi_{12} \\
p_{2|\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{\tau|1} &= 0 \\
p_{\tau|2} &= 0 \\
p_{\tau|\tau} &= \sigma
\end{aligned}$$

where $\Gamma = \lim_{\lambda_\tau \to 0} \frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}$.

Proof. In Appendix A.4 □

Note that when λ_τ goes to zero, we have $p_{\tau|1} = p_{\tau|2} = 0$, and individuals of type θ_1 and θ_2 are matched as if individuals θ_τ were not in the population. The conditional probabilities $p_{1|1}$, $p_{2|1}$, $p_{1|2}$ and $p_{2|2}$ are then consistent with the classical setting (Bergstrom, 2003; Alger and Weibull, 2013).

When the assortment matrix is uniformly constant, we have $\phi_{12} = \sigma$ and $\Gamma = 0$. The limit Γ can be interpreted as the marginal matching-probability difference between mutants and residents of the two types: $\Gamma = \lim_{\lambda_\tau \to 0}(p_{\tau 2} - p_{\tau 1})/\lambda_\tau$. In other words, if individuals θ_1 and θ_2 meet the mutants at the same rate when they enter the population, then $\Gamma = 0$, while if residents of one type meet the mutants at a higher rate than the other residents do then $\Gamma \neq 0$. Finally, when the assortment functions $\phi_{\tau 1}$ and $\phi_{\tau 2}$ are right-differentiable in $\lambda_\tau = 0$, we have $\Gamma = \partial \phi_{\tau 1}(\lambda, 0)/\partial \lambda_\tau - \partial \phi_{\tau 2}(\lambda, 0)/\partial \lambda_\tau$.[10] Therefore, Γ is the marginal assortment difference between mutants and residents of the two types.

2.2.3 Fitness game

The pairwise-matched individuals engage in a symmetric interaction.[11] Each individual is as likely to be in one or the other side of the interaction. We assume that the common strategy set X is a nonempty, compact and convex set in a topological vector space.[12] Following Güth and Yaari (1992), we adopt an indirect evolutionary framework. The behavior of individuals, i.e. the strategy they play, is driven by the maximization of personal preferences, which are described by a continuous utility function $u_{\theta_i} : X^2 \to \mathbb{R}$. On the other hand, the individuals'

[10] Because $\phi_{\tau 1}(\lambda, 0) = \phi_{\tau 2}(\lambda, 0) = \sigma$

[11] The framework can also be extended to asymmetric interactions with ex-ante symmetry.

[12] More precisely, we assume that X is a locally convex Hausdorff space. However, most of our analysis will focus on the simpler case of a finite two-player normal-form game where X is the set of mixed strategies.

evolutionary success is given by some exogenous payoff (fitness) function π, where we assume $\pi : X^2 \to \mathbb{R}$ to be continuous. The pair $< X, \pi >$ is called the *fitness game*.

To prevent individuals from deviating from their utility-maximization, we consider the individuals' preferences as their private information.[13] A Bayesian Nash Equilibrium (BNE) is then a set of strategies, one for each type, where each strategy is a best reply to the others in the given population state:

Definition 4 (Bayesian Nash Equilibrium)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, $(x_1, x_2, x_\tau) \in X^3$ is a type-homogeneous Bayesian Nash equilibrium if:

$$\forall i \in I: \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in I} p_{j|i} \cdot u_{\theta_i}(x, x_j) \tag{2.2}$$

The set of Bayesian Nash Equilibria in population state s, i.e. all solutions (x_1, x_2, x_τ) of (Eq. 2.2), is called $B^{NE}(s) \subseteq X^3$.

Remark 4. The definition of Bayesian Nash equilibrium remains valid when there is no mutant in the population, i.e. when the population is made of two types. In this case, (x_1, x_2) is a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \lambda)$ if for all $i \in \{1,2\}$, $x_i \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in \{1,2\}} p_{j|i} \cdot u_{\theta_i}(x, x_j)$.

Property 5. Since in the state $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ the residents are matched between them, as if there were no mutants in the population (Lemma 2), if $(x_1^\circ, x_2^\circ) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$, then for any strategy $x_\tau^\circ \in X$ such that $x_\tau^\circ \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in I} p_{j|\tau} \cdot u_{\theta_\tau}(x, x_j^\circ)$, we have $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$. Reciprocally, if $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, then $(x_1^\circ, x_2^\circ) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$.

We now define the equilibrium correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1)^2 \rightrightarrows X^3$. This correspondence maps the population share of each type to the associated equilibria. Using the definition of assortativity (Definition 3), it can be extended by continuity to $(0,1) \times [0,1)$ to cover the limit when the mutant share λ_τ goes to zero. The following lemma will be useful for the evolutionary stability analysis:

Lemma 3. *$B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.*
If for all $i \in I$ u_{θ_i} are concave in their first arguments, then $B^{NE}(s) \neq \emptyset$.
The correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous.

Proof. In Appendix B.1. □

[13] A large body of research has studied preference evolution under complete and incomplete information, showing that individuals adjust their behavior under complete information (e.g. Robson, 1990; Ellingsen, 1997; Bester and Güth, 1998; Possajennikov, 2000; Ok and Vega-Redondo, 2001; Sethi and Somanathan, 2001; Heifetz et al., 2007; Dekel et al., 2007). For example, suppose that two individuals are playing a prisoner's dilemma, where the first player prefers to defect and the second prefers to cooperate. Under incomplete information, each individual will stick to their original preference. But if the cooperator knows the preference of the defector, then she will deviate and also defect (See also Ockenfels, 1993, for a discussion of cooperation in prisoners' dilemma).

An individual of type θ_i who plays strategy $x_i \in X$ when her opponent of type θ_j plays strategy $x_j \in X$ gets material payoff $\pi(x_i, x_j)$. For simplicity, we will often note $\pi(x_i, x_j) \equiv \pi_{ij}$. For all $i \in I$, the fitness of a type θ_i is given by the average payoff obtained by individuals θ_i:

Definition 5 (Type fitness)**.** In a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, let $(x_1, x_2, x_\tau) \in B^{NE}(s)$. For all $i \in I$, the fitness of a type θ_i is given by:

$$\Pi_{\theta_i}(x_1, x_2, x_\tau, s) = \sum_{j \in I} p_{j|i} \cdot \pi(x_i, x_j) \tag{2.3}$$

2.2.4 Evolutionarily stable population

In order to analyze the evolutionary stability of a heterogeneous population, we need to extend the concept of *evolutionarily stable preference* (Alger and Weibull, 2013). An *evolutionarily stable population* should respect two conditions. First, the two resident types should earn the same type fitness to coexist. Second, the population must resist a small-scale invasion of any other type by earning a greater type fitness. Formally:

Definition 6 (Evolutionarily stable population)**.** A population in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ is evolutionarily stable against a mutant type $\theta_\tau \in \Theta$ such that for all $i \in \{1, 2\}$ $\theta_\tau \neq \theta_i$ if:

1. θ_1 and θ_2 earn the same type fitness: $\Pi_{\theta_1}(x_1^\circ, x_2^\circ, s^\circ) = \Pi_{\theta_2}(x_1^\circ, x_2^\circ, s^\circ)$ in all Bayesian Nash equilibria (x_1°, x_2°) in the population state s°;
2. θ_1 and θ_2 earn a greater type fitness than a small share of mutants: there exists an $\bar{\varepsilon} > 0$ such that for all $i \in \{1, 2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.

Moreover, a population is evolutionarily stable if it is evolutionarily stable against all types $\theta_\tau \in \Theta$ such that for all $i \in \{1, 2\}$, $\theta_\tau \neq \theta_i$.

The first condition of an *evolutionarily stable population* requires that the two residents earn the same type fitness. In the framework of evolutionary game dynamics, the evolution of strategies (and preferences) is dictated by an evolutionary process called a replicator, which usually depends on the difference between the fitness obtained and the average fitness in the population. If the fitness of a given type is greater than the average fitness, then the population share of this type will increase. Hence, the two resident types should get the same fitness for the population share λ° to be stable.

In the second condition defining an *evolutionarily stable population*, when the mutants enter the population, we allow the relative share of the two residents λ to change around a small neighborhood of its initial value λ°. However, in this case ($\lambda_\tau > 0$), we only impose that the two residents earn a greater type fitness than the mutant, and not that the two residents earn the same type fitness. Such a condition would be too restrictive. Thus, by entering the population, the mutant could destabilize the residents, i.e. one type could overcome (or invade) the other. To analyze if an *evolutionarily stable population* is robust to mutant entry, one would need to model the evolutionary dynamics. The results would then depend on the evolutionary process selected, which could be challenging in economics since this evolutionary process depends on genetic, cultural and technological transmission (Norton et al., 1998; Van Damme, 1991). Hence, such a analysis falls out of the scope of this study, but we will discuss the concept of robustness in more detail in Section 2.3.3.

The definition of *evolutionarily stable population* is consistent with the classical setting: an *evolutionarily stable preference* is an *evolutionarily stable population* when there is only one resident type and one mutant type. Moreover, this definition is similar to the concept of *evolutionarily stable configuration* by Dekel et al. (2007). A configuration (a distribution of preferences and the associated equilibria) is evolutionarily stable if it is balanced, i.e. if all types earn the same fitness, and if mutants do not outperform residents. Thus, an *evolutionarily stable population* can be understood as an *evolutionarily stable configuration* in which the distribution of preferences consists of the shares of each type. However, there are a few differences between the two definitions. First, the definition of *evolutionarily stable population* applies to preferences, and thus to all Bayesian Nash equilibria of the population. Second, by requiring that the mutant type is different from the residents in the definition of *evolutionarily stable population*, we can impose that resident individuals earn a strictly greater payoff than the mutants. Finally, the introduction of assortative matching limits the analysis to a finite number of types.

We will now derive two useful results linking the second condition of evolutionary stability with what is happening at the limit when the mutant share goes to zero. Recall that $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$ denotes a population state when the mutant share goes to zero.

Lemma 4. *When the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, if for all $i \in \{1,2\}$, $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) > \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ for all $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ then there exists $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. In Appendix B.2. □

Lemma 5. *When the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, if there exists $i \in \{1,2\}$ such that $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) < \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ with $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ a singleton, then there does not exist $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. In Appendix B.3. □

Lemmas 4 and 5 mean that it is generally sufficient to only study what is happening at the limit when the mutant share goes to zero when analyzing the evolutionary stability of a population. If the two residents earn the same type-fitness and a strictly greater playoff than any mutant $\theta_\tau \neq \theta_1, \theta_2$ in all Bayesian Nash equilibria in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, then the population $s^{\circ} = (\theta_1, \theta_2, \lambda^{\circ})$ is evolutionarily stable. Else, the population is generally not evolutionarily stable.[14] Note that the proof of Lemma 5 actually develops a stronger argument than "not evolutionarily stable". If the residents earn the same type fitness in $s = (\theta_1, \theta_2, \theta_\tau, \lambda^{\circ}, 0)$ and if the assumptions of Lemma 5 are satisfied, the proof shows that there exists an $\bar{\varepsilon} > 0$ such that the mutant earns a greater type fitness in all Bayesian Nash equilibria in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^{\circ}| < \bar{\varepsilon}$. Alger and Weibull (2013) call this property *evolutionary unstability.*

2.2.5 *Homo moralis*

In the classical setting with a homogeneous population, Alger and Weibull (2013) show that the only *evolutionarily stable preference* is the one of *homo hamiltonensis*, a particular kind of *homo moralis.*

Definition 7 (Homo moralis and homo hamiltonensis)**.** An individual is a *homo moralis* if her utility function is of the form:

$$u_\kappa(x, y) = (1-\kappa) \cdot \pi(x, y) + \kappa \cdot \pi(x, x) \tag{2.4}$$

where $\kappa \in [0, 1]$ is her degree of morality.
A *homo moralis* maximizes a convex combination of her classical selfish payoff, with a weight $(1-\kappa)$, and of her "moral" payoff, defined as the payoff she would get if her opponent plays like her, with a weight κ. If $\kappa = 0$, then the individual is a *homo oeconomicus* (fully selfish). If $\kappa = 1$, then the individual is a *homo kantiensis* (fully moral). If the degree of morality κ is equal to the assortativity σ, then the individual is called *homo hamiltonensis*[15].

In our analysis, we will often encounter *homo hamiltonensis.* More precisely the strategies played by *homo hamiltonensis* individuals when all residents are of this type, called *Hamiltonian strategies*, will play a key role in the analysis of evolutionary stability.

Definition 8 (Hamiltonian strategies)**.** $x_\sigma \in X$ is a *Hamiltonian strategy* if:

$$x_\sigma \in \underset{x \in X}{\text{argmax}} \quad u_\sigma(x, x_\sigma)$$

For all $y \in X$, we call $\beta_\sigma(y) = \text{argmax}_{x\in X}\ u_\sigma(x, y)$ the best-reply correspondence of *homo hamiltonensis* individuals, and we denote by $X_\sigma = \{x \in X : x \in \beta_\sigma(x)\}$ the set of fixed-points of *homo hamiltonensis.*

[14] The only undetermined cases are when the two residents earn the same type-fitness but (a) there exists a mutant θ_τ and a Bayesian Nash equilibra $(x_1^{\circ}, x_2^{\circ}, x_\tau^{\circ})$ of the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^{\circ}, 0)$ such that the residents and the mutant earn the same type-fitness: $\Pi_{\theta_i}(x_1^{\circ}, x_2^{\circ}, x_\tau^{\circ}, s) = \Pi_{\theta_\tau}(x_1^{\circ}, x_2^{\circ}, x_\tau^{\circ}, s)$; (b) we are in the case of Lemma 5 except that $B^{NE}(s^{\circ})$ is not a singleton and there also exists a Bayesian Nash equilibrium such that the residents earn a greater type fitness than the mutant.

[15] Alger and Weibull (2013) named *homo hamiltonensis* in homage to the late biologist William Donald Hamilton. See Grafen (2004) for a biography.

Consider a homogeneous population of *homo hamiltonensis* and a small group of mutants that wish to enter the population. If the mutant is not a "behavioral-alike"[16] to *homo hamiltonensis*, the mutant will always get a lower type fitness than *homo hamiltonensis*. For example, if the mutant is a *homo moralis* with a degree of morality different from the assortativity ($\kappa \neq \sigma$), such that this *homo moralis* and *homo hamiltonensis* are not behaviorally-alike, then to enter the population, the degree of morality of the *homo moralis* should evolve in direction of the assortativity.

But is this homogeneity a required feature of evolutionary stability? What happens when the population is more diverse? We explore these questions in this chapter, using a population of *homo oeconomicus* and *homo kantiensis* involved in a prisoners' dilemma as an illustration.

2.2.6 *Homo oeconomicus* and *homo kantiensis* in a prisoners' dilemma

A prisoners' dilemma is a finite symmetric fitness game with two pure strategies: cooperate (C) or defect (D). We denote π^{ij} the payoff obtained when pure strategy i is played against pure strategy j. A prisoners' dilemma is well defined when $\pi^{CD} < \pi^{DD} < \pi^{CC} < \pi^{DC}$. In other words, players benefit if they both cooperate instead of defecting ($\pi^{DD} < \pi^{CC}$), but each of them has an incentive to deviate ($\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$). In our analysis, the sum S_π will play an important role:

$$S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC} \tag{2.5}$$

Since $\pi^{CC} - \pi^{CD}$ is the gain minus the cost of cooperation and $\pi^{DC} - \pi^{DD}$ is the gain minus the cost of defection, S_π can be interpreted as the net benefit of cooperation minus the net benefit of defection. When $S_\pi = 0$ the game is sometimes called additive. Throughout this chapter, we will use three examples of the prisoners' dilemma: (a) $S_\pi < 0$, (b) $S_\pi = 0$ and (c) $S_\pi > 0$.

Table 2.1 – Prisoner's dilemma examples

(a)

	C	D
C	(4,4)	(0,6)
D	(6,0)	(1,1)

S_π=-1<0

(b)

	C	D
C	(4,4)	(0,5)
D	(5,0)	(1,1)

S_π=0

(c)

	C	D
C	(4,4)	(0,4.5)
D	(4.5,0)	(1,1)

S_π=0.5>0

Let A be the matrix of the payoffs in the game, $A = [\pi^{CC}, \pi^{CD}; \pi^{DC}, \pi^{DD}]$. We allow players to use mixed strategies so that the strategy set X is the segment $\Delta = \{z \in \mathbb{R}^2_+ : z_1 + z_2 = 1\}$, where z_1 the probability to cooperate and z_2 the probability to defect. The payoff obtained by an individual playing strategy $x_1 \in X = \Delta$ when matched with an individual playing $x_2 \in X$ is then:

[16] Types θ and τ are called behavioral-alike if they are behaviorally indistinguishable. Precisely, with θ being the resident, the set of of types τ that are behaviorally alike to θ is called Θ_θ:

$$\Theta_\theta = \{\tau \in \Theta : \exists x \in X_\theta \; s.t. (x,x) \in B^{NE}(\theta, \tau, 0)\}$$

$\pi(x_1,x_2) = x_1^\intercal A x_2$, where $\pi : X^2 \to \mathbb{R}$ is a bilinear function. Since X is a segment, individuals' decisions are fully characterized by their probability to cooperate. We will denote $\alpha_i \in [0,1]$ the probability of an individual of type θ_i to cooperate. Hence, the payoff obtained by an individual θ_1 playing strategy $x_1 \in X$ when matched with an individual θ_2 playing $x_2 \in X$ is:

$$\pi(x_1,x_2) = \alpha_1\alpha_2\pi^{CC} + \alpha_1(1-\alpha_2)\pi^{CD} + (1-\alpha_1)\alpha_2\pi^{DC} + (1-\alpha_1)(1-\alpha_2)\pi^{DD}$$

Individuals *homo oeconomicus* are fully selfish, their morality coefficient is $\kappa = 0$ so that their utility is $u_0(x,y) = \pi(x,y)$. Hence, they always defect in a prisoner's dilemma because $\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$. Formally, for all $(x,y) \in X^2$ with $x = (\alpha_x; 1-\alpha_x)$, $\alpha_x \neq 0$ (i.e. x is not defection) and $y = (\alpha_y; 1-\alpha_y)$, we have:

$$\begin{aligned}
u_0(D,y) - u_0(x,y) &= \left[\alpha_y\pi^{DC} + (1-\alpha_y)\pi^{DD}\right] \\
&\quad - \left[\alpha_x\alpha_y\pi^{CC} + \alpha_x(1-\alpha_y)\pi^{CD} + (1-\alpha_x)\alpha_y\pi^{DC} + (1-\alpha_x)(1-\alpha_y)\pi^{DD}\right] \\
&= \alpha_x\left[\alpha_y\left(\pi^{DC} - \pi^{CC}\right) + \left(1-\alpha_y\right)\left(\pi^{DD} - \pi^{CD}\right)\right] \\
&> 0
\end{aligned}$$

On the other hand, individuals *homo kantiensis* are fully moral, their morality coefficient is $\kappa = 1$ so that their utility is $u_1(x,y) = \pi(x,x)$. They always cooperate in a prisoner's dilemma because $\pi^{CC} > \pi^{DD}$. It is worth noting that the utility of a *homo kantiensis* individual does not depend on her opponent strategy but only on her own strategy.

Throughout this chapter, we will analyze the evolutionary stability of a population of *homo oeconomicus* (θ_1) and *homo kantiensis* (θ_2) in the state $s = (\theta_1,\theta_2,\lambda)$, with $\lambda \in (0,1)$ the share of *homo kantiensis*. Consequently, the only Bayesian Nash equilibrium in the population state s is $(x_1,x_2) = (D,C)$, or alternatively $(\alpha_1,\alpha_2) = (0,1)$. Moreover, the share of *homo kantiensis* λ is also equal to the cooperation share in the population.

2.3 Is a heterogeneous population of *homo oeconomicus* and *homo kantiensis* favored by evolution?

We consider a population of *homo oeconomicus* and *homo kantiensis* involved in a prisoner's dilemma. In section 2.3.1, we analyze when *homo oeconomicus* and *homo kantiensis* can coexist. In Section 2.3.2, we analyze the evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis*. Finally, in Section 2.3.3 we discuss the robustness of *evolutionarily stable populations*.

2.3.1 On the coexistence of *homo oeconomicus* and *homo kantiensis*

The first condition of evolutionary stability requires that the residents earn the same type fitness in all Bayesian Nash equilibria in the state s (Definition 6). In this section, we explore when this condition is satisfied. Let θ_1 be *homo oeconomicus*, θ_2 *homo kantiensis* and $\lambda^\circ \in (0,1)$ the share of *homo kantiensis*. The only Bayesian Nash equilibrium in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ is $(x_1, x_2) = (D, C)$ (see Section 2.2.6). Hence, *homo oeconomicus* and *homo kantiensis* earn the same type fitness if and only if:

$$\Pi_{\theta_1}(D,C,s^\circ) = \Pi_{\theta_2}(D,C,s^\circ) \tag{2.6}$$

Using Lemma 2 and noting $\phi_{12} \equiv \phi_{12}(\lambda^\circ, 0)$, we can write the type fitness of *homo oeconomicus* and *homo kantiensis* in function of the share λ° and of the assortment between *homo oeconomicus* and *homo kantiensis* when there is no mutant in the population:

$$\begin{aligned} \Pi_{\theta_1}(D,C,s^\circ) &= \left[(1-\lambda^\circ) + \lambda^\circ \cdot \phi_{12}\right] \cdot \pi^{DD} + \left[\lambda^\circ(1-\phi_{12})\right] \cdot \pi^{DC} \\ \Pi_{\theta_2}(D,C,s^\circ) &= \left[(1-\lambda^\circ)(1-\phi_{12})\right] \cdot \pi^{CD} + \left[\lambda^\circ + (1-\lambda^\circ)\phi_{12}\right] \cdot \pi^{CC} \end{aligned} \tag{2.7}$$

Consequently, noting $\Pi_{\theta_{1-2}} \equiv \Pi_{\theta_1}(D,C,s^\circ) - \Pi_{\theta_2}(D,C,s^\circ)$ we have:

$$\Pi_{\theta_{1-2}} = \left[\pi^{DD} - \pi^{CD} - \phi_{12}\left(\pi^{CC} - \pi^{CD}\right)\right] - \lambda^\circ\left(1-\phi_{12}\right)\left[\pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}\right] \tag{2.8}$$

Similarly:

$$\Pi_{\theta_{1-2}} = (1-\lambda^\circ)\left(1-\phi_{12}\right)\left[\pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}\right] - \left[\pi^{CC} - \pi^{DC} - \phi_{12}\left(\pi^{DD} - \pi^{DC}\right)\right] \tag{2.9}$$

We define: $Q_\pi \equiv \pi^{DD} - \pi^{CD} - \phi_{12}(\pi^{CC} - \pi^{CD})$ and $R_\pi \equiv \pi^{CC} - \pi^{DC} - \phi_{12}(\pi^{DD} - \pi^{DC})$. Note that we have: $Q_\pi + R_\pi = (1-\phi_{12})S_\pi$, with $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$. Rewriting the type-fitness equality (Equation 2.6) with Equations 2.8 and 2.9, we obtain two equivalent conditions, one for λ° and the other for $(1-\lambda^\circ)$:

$$\begin{aligned} \lambda^\circ\left(1-\phi_{12}\right)S_\pi &= Q_\pi \\ \left(1-\lambda^\circ\right)\left(1-\phi_{12}\right)S_\pi &= R_\pi \end{aligned}$$

We have the following proposition:

Proposition 2 (Type-fitness equality). *In the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness if and only if:*

1. *When $S_\pi = 0$: $Q_\pi = 0$, i.e. $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.*
2. *When $S_\pi \neq 0$: $\lambda^\circ = Q_\pi / \left[\left(1-\phi_{12}\right)S_\pi\right]$.*

Moreover, if homo oeconomicus and homo kantiensis earn the same type fitness, then $\phi_{12} \in (0,1)$.

Proof. In Appendix B.4. □

Proposition 2 characterizes the conditions under which *homo oeconomicus* and *homo kantiensis* can coexist in any prisoners' dilemma. In other words, the proposition provides information on the existence of a population of *homo oeconomicus* and *homo kantiensis* earning the same type fitness. If there exists $\lambda^\circ \in (0,1)$ such that $\phi_{12} = (\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD})$ when $S_\pi = 0$ or $\lambda^\circ = Q_\pi / \left[\left(1 - \phi_{12}\right) S_\pi\right]$ when $S_\pi \neq 0$, then *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$.

Although our analysis is static, there is a link between Proposition 2 and the evolutionary game dynamics framework. Indeed, at the equilibrium in a dynamic game, the two types should earn the same fitness. Thus, Proposition 2 allows to quickly identify the candidate population-state for an equilibrium in a dynamic game. The remaining question in this context is then whether or not this equilibrium can be reached. The answer depends not only on the replicator but also on the shape of the assortment function.

Finally, the last part of the Proposition stipulates that the assortment should be in a given range $\phi_{12} \in (0,1)$ to allow *homo oeconomicus* and *homo kantiensis* earning the same type fitness. This range is detailed in the proof of the Proposition:

1. When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD}) < \phi_{12} < (\pi^{DC} - \pi^{CC}) / (\pi^{DC} - \pi^{DD})$.
2. When $S_\pi = 0$: $\phi_{12} = (\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD})$.
3. When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC}) / (\pi^{DC} - \pi^{DD}) < \phi_{12} < (\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD})$.

Thus, assortative matching plays a key role in allowing a heterogeneous population of *homo oeconomicus* and *homo kantiensis*. In other words, assortment is critical to better understand cooperative behaviors, as already pointed out by Eshel and Cavalli-Sforza (1982), Bergstrom (2003) or Allen and Nowak (2015) among others.[17] We will further discuss this result looking at the case of a uniformly-constant assortment.

Coexistence under uniformly-constant assortment

We now consider the case of a uniformly-constant assortment (Definition 2), which is an extension of uniform random matching accounting for assortatively-matched interactions. Under uniformly-constant assortment, the assortment functions are constant and equal to the assortativity σ (Definition 3) by continuity: for all $\lambda \in (0,1)$, $\phi_{12}(\lambda, 0) = \sigma \in [0,1]$.

The following Corollary recaps the results of Proposition 2 under uniformly-constant assortment:

[17] Cooperative behaviors can also arise thanks to reciprocity and punishment (see e.g. Fehr and Gächter, 2002; Bowles and Gintis, 2004; Nowak and Sigmund, 2005) and when participation in a public good game is optional (Hauert et al., 2002).

Corollary 1 (Type-fitness equality under uniformly-constant assortment)**.** *In the population state $s = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness under uniformly-constant assortment if and only if:*

1. *When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD}) < \sigma < (\pi^{DC} - \pi^{CC}) / (\pi^{DC} - \pi^{DD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma)S_\pi]$.*
2. *When $S_\pi = 0$: $\sigma = (\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD})$.*
3. *When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC}) / (\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD}) / (\pi^{CC} - \pi^{CD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma)S_\pi]$.*

Proof. In Appendix B.4. □

There exists a population share $\lambda^\circ \in (0,1)$ such that *homo oeconomicus* and *homo kantiensis* earn the same type fitness if the assortativity is in a given range. This result is quite intuitive. Indeed, if the assortment is too low then *homo oeconomicus* earns a greater type fitness than *homo kantiensis*. For instance under uniform random matching (for all $\lambda \in (0,1)$, $\phi_{12} = \sigma = 0$), we have:

$$\Pi_{\theta_1}(D,C,s^\circ) = (1-\lambda^\circ)\pi^{DD} + \lambda^\circ \pi^{DC}$$
$$\Pi_{\theta_2}(D,C,s^\circ) = (1-\lambda^\circ)\pi^{CD} + \lambda^\circ \pi^{CC}$$

Since $\pi^{CD} < \pi^{DD}$ and $\pi^{CC} < \pi^{DC}$, $\Pi_{\theta_1}(D,C,s^\circ) > \Pi_{\theta_2}(D,C,s^\circ)$. Conversely, if the assortment is too high then *homo kantiensis* earns a greater type-fitness than *homo oeconomicus*. For instance, let $\sigma = 1$. This means that *homo oeconomicus* and *homo kantiensis* individuals only meet individuals of their own type. Thus, we have $\Pi_{\theta_1}(D,C,s^\circ) = \pi^{DD}$, and $\Pi_{\theta_2}(D,C,s^\circ) = \pi^{CC}$. Since $\pi^{CC} > \pi^{DD}$, $\Pi_{\theta_1}(D,C,s^\circ) < \Pi_{\theta_2}(D,C,s^\circ)$.

Note that when $S_\pi = 0$, i.e. when the the game is additive, there is a unique assortativity σ allowing *homo oeconomicus* and *homo kantiensis* to earn the same type-fitness. When the assortativity is below this threshold, *homo oeconomicus* dominates, while *homo kantiensis* dominates when the assortativity is above this threshold. This result is in line with the literature. For instance, Bergstrom (2003) and Allen and Nowak (2015) have studied the evolution of cooperative strategies in an evolutionary game dynamics framework, finding that assortment allows cooperation in prisoner's dilemma. Since at the equilibrium, strategies must earn the same fitness, their results are consistent with ours. In particular, in a simplified version of the game, where payoffs are additive ($\pi^{CD} = -c$, $\pi^{DD} = 0$, $\pi^{CC} = b - c$ and $\pi^{DC} = b$ with $b > c > 0$, $S_\pi = 0$) and the assortment constant, they highlight that cooperation is favored when a condition similar to the Hamilton's rule is satisfied.[18] We obtain an analogous condition in this simplified game: cooperation will outperform defection when $b\sigma > c$.

[18] Hamilton's rule stipulates that the frequency of an altruistic gene will increase if $br > c$, with b the reproductive gain for the recipient of the altruistic act, c the reproductive cost for the altruist individual, and r the genetic relatedness of the recipient to the actor (Hamilton, 1964a,b).

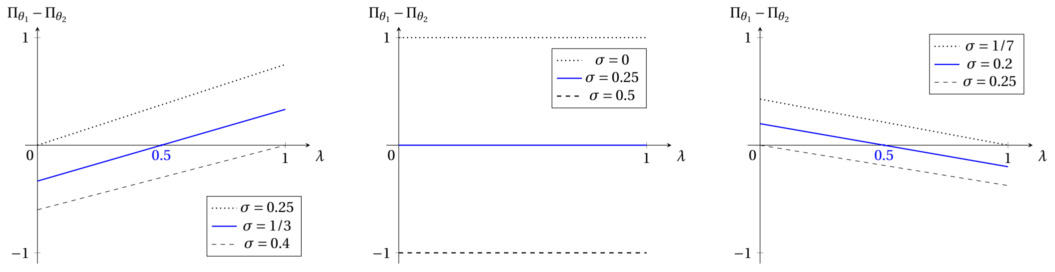
We now illustrate Corollary 1 with the examples defined in Section 2.2.6.

(a) First, let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 6$. We then have $S_\pi = -1 < 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -2 + 5\sigma$. Thus, there exists a heterogeneous population satisfying type-fitness equality when $0.25 < \sigma < 0.4$ (see Figure 2.2a). With $\sigma = 1/3$, then $\lambda^\circ = 0.5$ and *homo kantiensis* and *homo oeconomicus* co-exist and get the same type fitness equal to $\Pi_\theta = 8/3$. If the assortment is too low ($\sigma \leq 0.25$), only *homo oeconomicus* survives. In contrast, when the assortment is too high ($\sigma \geq 0.4$), *homo kantiensis* would dominate.

(b) Now let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 5$. We have $S_\pi = 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -1 + 4\sigma$. Thus, the only assortativity value consistent with type-fitness equality is $\sigma = 0.25$ (see Figure 2.2b). But then, for any population share $\lambda^\circ \in (0, 1)$, *homo kantiensis* and *homo oeconomicus* earn the same type-fitness.

(c) Finally, let $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$ and $\pi^{DC} = 4.5$. We have $S_\pi = 0.5 > 0$, $Q_\pi = 1 - 4\sigma$ and $R_\pi = -0.5 + 3.5\sigma$. Thus, there exists a heterogeneous population satisfying type-fitness equality when $1/7 < \sigma < 0.25$ (see Figure 2.2c). For example, when $\sigma = 0.2$, then $\lambda^\circ = 0.5$ and *homo kantiensis* and *homo oeconomicus* live together and get the same type-fitness equal to $\Pi = 2.4$. As above, the assortment plays a key role: if too low or too high, one type will dominate.

The assortativity allowing a heterogeneous population when $S_\pi = 0$ is $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) = 0.25$. It is also the minimum assortativity for a heterogeneous population when $S_\pi < 0$ and the maximum assortativity for a heterogeneous population when $S_\pi > 0$. This comes as no surprise. Indeed, as discussed in Section 2.2.6, S_π can be interpreted as the net benefit of cooperation minus the net benefit of defection. Hence, when $S_\pi < 0$, defectors (*homo oeconomicus*) have an advantage and only high values of assortativity allows a heterogeneous population. Reciprocally, when $S_\pi > 0$, the game favors cooperators (*homo kantiensis*) and a lower value of assortativity is needed to get a heterogeneous population.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 2.2 – Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under uniformly-constant assortment

Coexistence under state-dependent assortment

As highlighted in the literature, the phenomenon of homophily is highly dependent on the context. The size and demographic characteristics of the community considered affect the degree of homophily among its members (McPherson et al., 2001; Currarini et al., 2009).[19] Therefore, going beyond the case of uniformly-constant[20] assortment, we pursue our analysis with the general case of a state-dependent assortment.

For this purpose, we define the function $\Pi_{\theta_{1-2}} : (0,1) \to \mathbb{R}$ as the type-fitness difference between *homo oeconomicus* and *homo kantiensis*. From Equation 2.8, we have for all $\lambda \in (0,1)$:

$$\Pi_{\theta_{1-2}}(\lambda) = Q_\pi(\lambda) - \lambda\left(1 - \phi_{12}(\lambda)\right) S_\pi$$

Where $\phi_{12}(\lambda) \equiv \phi_{12}(\lambda, 0)$ and $Q_\pi(\lambda) \equiv \pi^{DD} - \pi^{CD} - \phi_{12}(\lambda)\left(\pi^{CC} - \pi^{CD}\right)$. By assumption, the assortment function is continuous in λ. Moreover, the examples considered in this section converge when λ goes to zero (i.e. *homo kantiensis* is a mutant) and when λ goes to one (i.e. *homo oeconomicus* is a mutant). Thus, the function $\Pi_{\theta_{1-2}}$ can be extended by continuity to $[0,1]$.

Given the great number of cases offered by the relaxation of the uniformly-constant assortment hypothesis, we consider three specific cases to illustrate Proposition 2:

1. In the first case, we suppose that ϕ_{12} is linear: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$ (see Figure 2.3). Thus, when the share of *homo kantiensis* λ goes to zero, $\phi_{12}(0) = 0.32$. This means that when *homo kantiensis* is a mutant, the probability for a *homo kantiensis* individual to meet another *homo kantiensis* is $p_{2|2} = 0.32$ (see Lemma 2). Recyprocally, when the share of *homo kantiensis* λ goes to one, $\phi_{12}(1) = 0.08$ so that the probability for a *homo oeconomicus* individual to meet another *homo oeconomicus* is $p_{1|1} = 0.08$. Hence, the shape of $\phi_{12}(\cdot)$ increases the evolutionary-success opportunities of each type: a *homo oeconomicus* is better off when its probability to meet another *homo oeconomicus* is low, while a *homo kantiensis* is better off meeting another *homo kantiensis* with a high probability.
2. In the second case, we suppose that ϕ_{12} is a U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$ (see Figure 2.3). With this shape, there is a high assortment when the population is imbalanced (i.e. when one resident accounts for a high share of the population), and the assortment is lower when the population is more balanced. This could represent a population where individuals are living nearby each other when their share in the population is low (or in other words, mutants enter the population in a specific area) while individuals are more mixed when the population is more balanced.
3. In the third case, we suppose that ϕ_{12} is an inverse U-shaped parabola: for all $\lambda \in [0,1]$,

[19] More precisely, Currarini et al. (2009) find that the *homophily* in most US ethnic groups is nonlinear and non-monotonous in the group size and McPherson et al. (2001) shows that *homophily* depends on sociodemographic, behavioral, and intrapersonal characteristics.

[20] Recall that under uniformly-constant assortment, the assortative matching is uniform across all types in the population and independent of the shares in the population.

$\phi_{12}(\lambda) = 2\lambda(1-\lambda)$ (see Figure 2.3). With this shape, the assortment is higher for a more balanced population. Bergstrom (2003) has shown that in a prisoners' dilemma involving cooperators and defectors, the assortment could have this shape when players have some choice about their partners. Moreover, such an assortment function is consistent with empirical evidence on the homophiliy in US ethnic groups (Currarini et al., 2009).

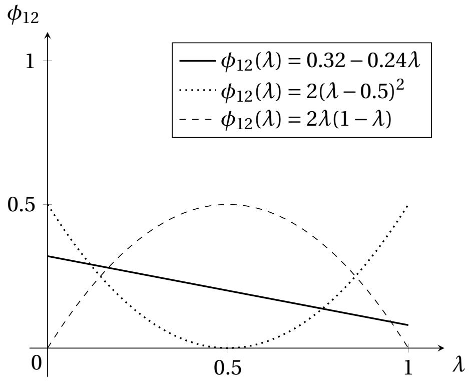


Figure 2.3 – Illustrative state-dependent assortment functions between *homo oeconomicus* and *homo kantiensis*

For each case, we consider the same examples studied above and defined in Section 2.2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$. Thus, $Q_\pi(\lambda) = 1 - 4\phi_{12}(\lambda)$ and $S_\pi = 5 - \pi^{DC}$.

1. When $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$, $\Pi_{\theta_{1-2}}(\lambda) = -0.24 S_\pi \lambda^2 + (0.96 - 0.68 S_\pi)\lambda - 0.28$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has one root $\lambda^\circ \in (0,1)$: $\lambda^\circ = 1/6$ and then $\phi_{12}(\lambda^\circ) = 0.28$ (See Figure 2.4a).
   (b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a line which intersects the x-axis for $\lambda^\circ = 7/24 \in (0,1)$, and then $\phi_{12}(\lambda^\circ) = 0.25$ (See Figure 2.4b).
   (c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has one root $\lambda^\circ \in (0,1)$: $\lambda^\circ = 0.5$ and then $\phi_{12}(\lambda^\circ) = 0.2$ (See Figure 2.4c).

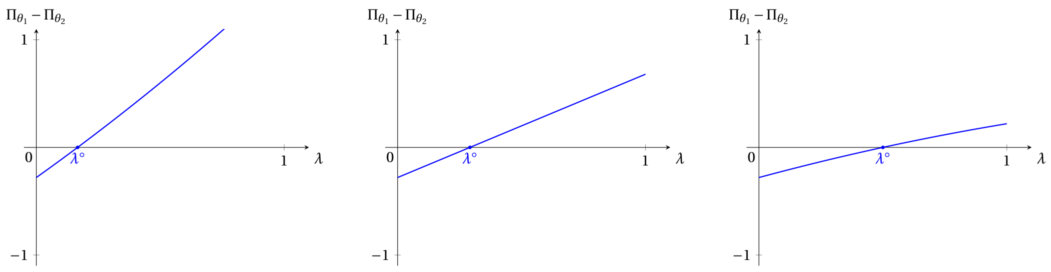
2. When $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$, $\Pi_{\theta_{1-2}}(\lambda) = 2S_\pi\lambda^3 - (2S_\pi + 8)\lambda^2 + (8 - 0.5S\pi)\lambda - 1$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.130$ and $\lambda^{\circ+} \approx 0.943$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.274$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.393$ (See Figure 2.5a).
   (b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has two roots in $(0,1)$: $\lambda^{\circ-} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ+} = 0.5 + 0.25\sqrt{2}$, and then $\phi_{12}(\lambda^{\circ-}) = \phi_{12}(\lambda^{\circ+}) = 0.25$ (See Figure 2.5b).
   (c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.157$ and $\lambda^{\circ+} \approx 790$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.235$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.168$ (See Figure 2.5c).

3. When $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$, $\Pi_{\theta_{1-2}}(\lambda) = -2S_\pi\lambda^3 + (2S_\pi + 8)\lambda^2 - (8 + S\pi)\lambda + 1$.
   (a) $S_\pi = -1 < 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ-} \approx 0.169$ and $\lambda^{\circ+} \approx 0.756$, and then $\phi_{12}(\lambda^{\circ-}) \approx 0.280$ and $\phi_{12}(\lambda^{\circ+}) \approx 0.369$ (See Figure 2.6a).

(b) $S_\pi = 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 2 which has two roots in $(0,1)$: $\lambda^{\circ -} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ +} = 0.5 + 0.25\sqrt{2}$, and then $\phi_{12}(\lambda^{\circ -}) = \phi_{12}(\lambda^{\circ +}) = 0.25$ (See Figure 2.6b). This case is actually symmetric to 2.(b) so that the equilibrium cooperation shares are the same.

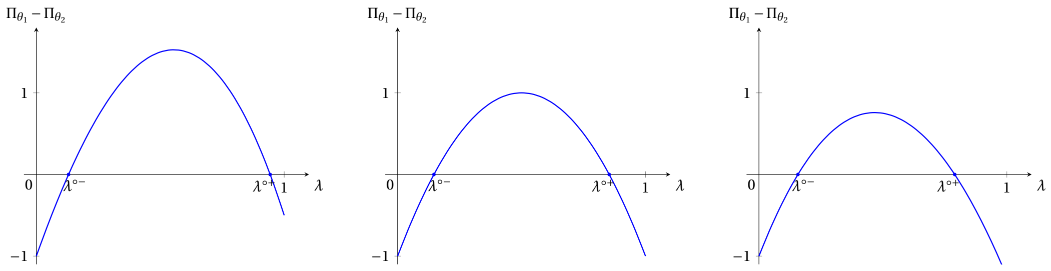
(c) $S_\pi = 0.5 > 0$, $\Pi_{\theta_{1-2}}$ is a polynom of degree 3 which has two roots in $(0,1)$: $\lambda^{\circ -} \approx 0.137$ and $\lambda^{\circ +} \approx 0.917$, and then $\phi_{12}(\lambda^{\circ -}) \approx 0.237$ and $\phi_{12}(\lambda^{\circ +}) \approx 0.153$ (See Figure 2.6c).

In each game, we find one cooperation share λ° allowing for a heterogeneous population with linear assortment (case 1) and two equilibrium cooperation shares with quadratic assortment (cases 2 and 3). However, this is not a general property of linear and quadratic assortment. The number of cooperation shares satisfying type-fitness equality depends on the game payoffs and on the assortment functions. Moreover, under linear assortment, note that the equilibrium cooperation share increases with S_π. Nonetheless, this is also not a general feature. For instance, with $\phi_{12}(\lambda) = 0.2\lambda + 0.2$, the equilibrium cooperation share decreases with S_π.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

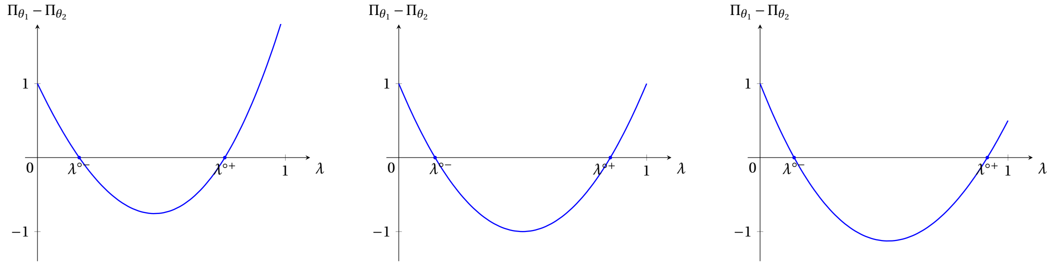
Figure 2.4 – Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 2.5 – Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$

State-dependent assortment brings more complexity but also more interesting equilibria. It will play a key role in the evolutionary-stability analysis (Section 2.3.2). Furthermore, the shape of the assortment function determines if an equilibrium cooperation share can be reached or not. We will discuss the dynamics in more detail in Section 2.3.3 on the robustness of *evolutionarily stable populations*.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 2.6 – Type-fitness difference in prisoner's dilemma between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) under state-dependent assortment $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$

2.3.2 On the evolutionary stability of heterogeneous populations

An *evolutionarily stable population* satisfies two conditions: residents earn the same type fitness and they resist a small-scale invasion of any other type (Definition 6). In the previous section, we studied when the first condition is met for a population of *homo oeconomicus* (θ_1) and *homo kantiensis* (θ_2). We now turn our analysis to the second condition, assuming that the residents earn the same type fitness in the Bayesian Nash equilibrium $(x_1, x_2) = (D, C)$ in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$, with $\lambda^\circ \in (0,1)$.

As shown in Lemmas 4 and 5, it is generally sufficient to only study what is happening at the limit when the mutant share goes to zero when analyzing the evolutionary stability of a population. Let $\theta_\tau \in \Theta$ be a mutant and (x_1, x_2, x_τ) a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Note that since *homo oeconomicus* individuals always defect no matter their opponent strategies while *homo kantiensis* individuals always cooperate, we have $(x_1, x_2, x_\tau) = (D, C, x_\tau)$. Using Lemma 2 and noting $\pi_{ij} \equiv \pi(x_i, x_j)$ and $\Pi_{\theta_i} \equiv \Pi_{\theta_i}(x_1, x_2, x_\tau, s)$ for all $(i, j) \in I^2$, we can write the type fitness of each type:

$$\begin{cases} \Pi_{\theta_1} = (1 - \lambda^\circ + \lambda^\circ \phi_{12}) \cdot \pi_{11} + \lambda^\circ (1 - \phi_{12}) \cdot \pi_{12} \\ \Pi_{\theta_2} = (1 - \lambda^\circ)(1 - \phi_{12}) \cdot \pi_{21} + [\lambda + (1 - \lambda)\phi_{12}] \cdot \pi_{22} \\ \Pi_{\theta_\tau} = [(1 - \lambda^\circ)(1 - \sigma) - \lambda^\circ (1 - \lambda^\circ)\Gamma] \cdot \pi_{\tau 1} + [\lambda^\circ (1 - \sigma) + \lambda^\circ (1 - \lambda^\circ)\Gamma] \cdot \pi_{\tau 2} + \sigma \cdot \pi_{\tau\tau} \end{cases}$$

Note that $\pi_{1\tau}$ and $\pi_{2\tau}$ do not appear in the expression of the type fitness of *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2}) because at the limit when the mutant share goes to zero, the residents are matched between them as if there were no mutants in the population. Consequently, since by assumption *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the Bayesian Nash equilibrium (D, C) in the population state s°, they also earn the same type fitness in all Bayesian Nash equilibria in the population state s, i.e. $\Pi_{\theta_1} = \Pi_{\theta_2} \equiv \Pi_\theta$.

Next, since we are in a two-strategies game, we can express the strategy x_τ in function of the strategies x_1 and x_2. For this purpose, recall that for all $i \in I$, $\alpha_i \in [0,1]$ is the probability that θ_i individuals attach to cooperation, so that $x_i = (\alpha_i; 1 - \alpha_i)$. When $\alpha_1 \neq \alpha_2$, there exists $\gamma \in \mathbb{R}$

such that $\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2$. The following Lemma depicts the difference in type-fitness between the residents and any mutant:

Lemma 6 (Difference in type fitness between residents and mutant)**.** *Let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$*, with* $\lambda^\circ \in (0,1)$*, engaged in a prisoners' dilemma such that the residents earn the same type fitness* Π_θ *for* $(x_1, x_2) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$ *with* $x^1 \neq x^2$*. Then, the difference in type-fitness between the residents and the mutant for* $(x_1, x_2, x_\tau) \in B^{NE}(s)$ *is:*

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} = {} & [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot S_\pi \\ & + [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{CC}-\pi^{CD}) + (1-\alpha_2)(\pi^{DC}-\pi^{DD})]\end{aligned}$$

Proof. In Appendix B.5. □

Since *homo oeconomicus* individuals defect and *homo kantiensis* individuals cooperate, we have $\alpha_1 = 0$, $\alpha_2 = 1$ and $\gamma = \alpha_\tau$. We now have all the ingredients to examine the evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis.* As in the coexistence analysis, we start with the case of uniformly-constant assortment before looking at the case of state-dependent assortment.

Evolutionary stability under uniformly-constant assortment

Under uniformly-constant assortment, we have $\phi_{12} = \sigma$ (by definition, see Remark 2) and $\Gamma = 0$. Indeed, $\Gamma = \lim_{\lambda_\tau \to 0}(\phi_{\tau 1} - \phi_{\tau 2})/\lambda_\tau$, and $\phi_{\tau 1} = \phi_{\tau 2} = \sigma$. As discussed in Section 2.2.2, Γ can be interpreted as the marginal matching-probability difference between mutants and residents of the two types. When individuals θ_1 and θ_2 meet the mutants at the same rate when they enter the population, then $\Gamma = 0$. We can rewrite Lemma 6 for the case of uniformly-constant assortment:

Corollary 2 (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis and* $\lambda^\circ \in (0,1)$*, engaged in a prisoners' dilemma such that the residents earn the same type fitness* Π_θ *for* $(D, C) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$*. Then, the difference in type-fitness between the residents and the mutant for* $(D, C, x_\tau) \in B^{NE}(s)$ *is:*

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. In Appendix B.5. □

This expression is much simpler than the general case. The difference in type-fitness between the residents and the mutant depends only on the assortativity, on the mutant's strategy and on the net benefit of cooperation minus the net benefit of defection S_π. Moreover, from Corollary 1, we know that $\sigma > 0$ because we assumed that *homo oeconomicus* and *homo kantiensis* were earning the same type fitness. Note also that $\alpha_\tau(1-\alpha_\tau) \geq 0$ because $(\alpha_\tau) \in [0,1]$ and if mutants do not play a pure strategy, the inequality is strict, i.e. $\alpha_\tau(1-\alpha_\tau) > 0$. Hence, the sign of the difference in type-fitness depends only on the sign of S_π.

Interestingly, the same expression remains valid in a more general case, when the mutant share is not equal to zero:

Lemma 7 (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis, engaged in a prisoners' dilemma. Then, we have for any* $(D, C, x_\tau) \in B^{NE}(s)$*:*

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau \Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. In Appendix B.6. □

Before stating our main results, we need to introduce two additional notions. First, the type set Θ is called *rich* if for each strategy $x \in X$, there exists some type $\theta \in \Theta$ for which this strategy is strictly dominant: $u_\theta(x,y) > u_\theta(x',y)$ for all $x' \neq x$ and y in X. When Θ is *rich*, for any strategy $x \in X$ it is always possible to find a mutant playing x. Second, we call Θ_{12} the set of mutants τ that are behaviorally indistinguishable from residents θ_1 and θ_2:

$$\Theta_{12} = \left\{\theta_\tau \in \Theta : \exists\, x \in X \,\text{such that}\, (x_1, x, x) \,\text{or}\, (x, x_2, x) \in B^{NE}(s)\right\}$$

In our study, the set Θ_{12} includes all the mutants that cooperate or defect when their share goes to zero, i.e. $\Theta_{12} = \left\{\theta_\tau \in \Theta : (D,C,D) \,\text{or}\, (D,C,C) \in B^{NE}(s)\right\}$. We have the following Theorem:

Theorem 1 (Evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis*)**.** *In a prisoners' dilemma under uniformly-constant assortment when* Θ *is rich, there exists a heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis against all types* $\theta_\tau \notin \Theta_{12}$ *if and only if* $S_\pi > 0$ *and* $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.

Moreover, if there exists a heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis, then it is unique and the cooperation share satisfies $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$.

Proof. In Appendix B.7. □

Theorem 1 fully characterizes the existence and uniqueness of a *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* in prisoners' dilemmas under uniformly-constant assortment. In particular, there does not exist a heterogeneous *evolutionarily stable population* when $S_\pi \leq 0$. When $S_\pi > 0$, there exists a unique heterogeneous *evolutionarily stable population* when the assortativity belongs to a range such that *homo oeconomicus* and *homo kantiensis* can coexist.

We made a few assumptions to derive Theorem 1. First, we assumed that the type set Θ was rich. If it was not, *homo oeconomicus* and *homo kantiensis* could be the only types in Θ. Then, any heterogeneous population satisfying type-fitness equality would be evolutionarily stable, even when $S_\pi \leq 0$ (because there does not exist any mutant). We could actually relax this assumption by assuming that there exists one type $\theta_\tau \in \Theta$ committed to a mixed strategy. Indeed, when $S_\pi \leq 0$, any mixed strategy enables the mutant to earn a greater type fitness than at least one of the residents. Second, the population is evolutionarily stable against all types $\theta_\tau \notin \Theta_{12}$, i.e. the types which are not behaviorally-alike to the residents. Indeed, if mutants cooperate or defect, then the share of cooperation changes and the mutant cannot earn a strictly smaller type fitness in all Bayesian Nash equilibria in a neighborhood of $\lambda°$.

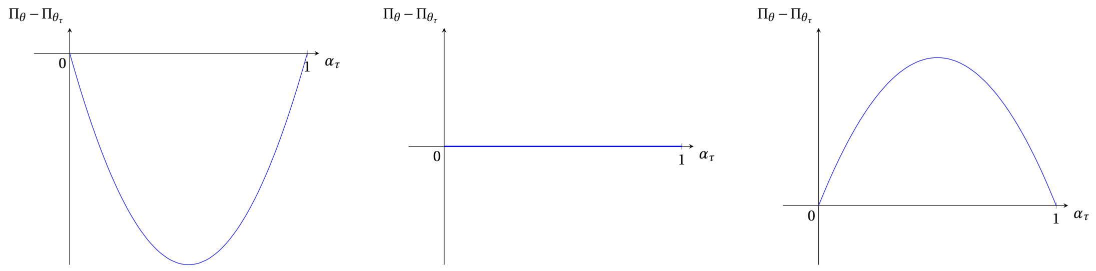
We now illustrate the Theorem going back to the examples defined in Section 2.2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$.

(a) First, $S_\pi < 0$. With a uniformly-constant assortment $\sigma = 1/3$, then with $\lambda° = 0.5$ the population satisfies type-fitness equality and $\Pi_\theta = 8/3$ (see Section 2.3.1). However, we have $S_\pi = -1$, and since the difference in type fitness between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$ (Corollary 2), any mutant would earn more than the residents at the limit as illustrated in Figure 2.7a. Hence we can conclude that the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable.

(b) Second, $S_\pi = 0$ and the game is additive. As discussed in Section 2.3.1, the only uniformly-constant assortment allowing type-fitness equality is $\sigma = 0.25$. With this value, for any $\lambda° \in (0,1)$ *homo oeconomicus* and *homo kantiensis* earns the same type fitness. On the other hand, any mutant would also earn the same type-fitness at the limit (see Figure 2.7b). From Lemma 7, the mutant would also earn a greater type-fitness than at least one of the residents when its share λ_τ increases. Thus the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable.

(c) Finally, $S_\pi > 0$. With a uniformly-constant assortment $\sigma = 0.2$, then with $\lambda° = 0.5$ the population satisfies type-fitness equality and $\Pi_\theta = 2.4$ (see Section 2.3.1). Moreover, we have $S_\pi = 0.5$, and the difference in type fitness between the residents and the mutant at the limit is: $\Pi_\theta - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)S_\pi$ (Corollary 2). Thus, for all $\alpha_\tau \in (0,1)$, $\Pi_\theta - \Pi_\tau > 0$ (see Figure 2.7c) and we can conclude that the population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable against all mutants which do not cooperate or defect.

Under uniformly-constant assortment, we can establish a link between evolutionary stability of heterogeneous and homogeneous populations. The only *evolutionarily stable preference* in a homogeneous population is *homo hamiltonensis*, a *homo moralis* with a degree of morality equal to the assortativity (Alger and Weibull, 2013). When *homo hamiltonensis* is the only resident, individuals of this type play *Hamiltonian strategies* $x_\sigma \in X_\sigma$ (see Definition 8). It turns out that in a heterogeneous *evolutionarily stable population* under uniformly-constant assortment, *homo oeconomicus* and *homo kantiensis* also play *Hamiltonian strategies*. The following Lemma details the *Hamiltonian strategies* in a prisoners' dilemma. Note that the set



(a) Non evolutionarily-stable ($S_\pi < 0$, $\sigma = 1/3$, $\lambda^\circ = 0.5$)

(b) Non evolutionarily-stable ($S_\pi = 0$, $\sigma = 0.25$)

(c) Evolutionarily stable ($S_\pi > 0$, $\sigma = 0.2$, $\lambda^\circ = 0.5$)

Figure 2.7 – Type-fitness difference between a heterogeneous population of *homo oeconomicus* and *homo kantiensis* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under uniformly-constant assortment.

of *Hamiltonian strategies* X_σ is expressed in function of the probability to cooperate. In other words, if $0 \in X_\sigma$ then defection is a *Hamiltonian strategy* (probability zero to cooperate) while if $1 \in X_\sigma$ then cooperation is a *Hamiltonian strategy.*

Lemma 8 (*Homo hamiltonensis* behavior in prisoners' dilemma)**.** *Let* $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$, $Q_\pi \equiv \pi^{DD} - \sigma\pi^{CC} - (1-\sigma)\pi^{CD}$ *and* $R_\pi \equiv \pi^{CC} - \sigma\pi^{DD} - (1-\sigma)\pi^{DC}$.
When $\sigma = 0$*, homo hamiltonensis is homo oeconomicus and always defects:* $X_\sigma = \{0\}$.
When $\sigma > 0$,

1. *If* $S_\pi < 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \text{if} \quad R_\pi \leq S_\pi \\ \left\{\frac{S_\pi - R_\pi}{(1+\sigma)S_\pi}\right\}, & \text{if} \quad R_\pi > S_\pi \quad \text{and} \quad Q_\pi > S_\pi \\ \{1\}, & \text{if} \quad Q_\pi \leq S_\pi \end{cases}$$

2. *If* $S_\pi = 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \text{if} \quad R_\pi < S_\pi \\ [0,1], & \text{if} \quad R_\pi = S_\pi \\ \{1\}, & \text{if} \quad R_\pi > S_\pi \end{cases}$$

3. *If* $S_\pi > 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \text{if} \quad R_\pi < 0 \\ \{0,1\}, & \text{if} \quad Q_\pi, R_\pi \geq 0 \\ \{1\}, & \text{if} \quad Q_\pi < 0 \end{cases}$$

Proof. In Appendix B.8. □

When there exists an *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis*, we have Q_π, R_π $S_\pi > 0$ (Theorem 1). Hence the *Hamiltonian strategies* are defection and cooperation (Lemma 8) confirming that *homo oeconomicus* and *homo kantiensis* play *Hamiltonian strategies*. When $Q_\pi = R_\pi = S_\pi = 0$, $X_\sigma = [0,1]$, i.e. all strategies are *Hamiltonian strategies*. This prevents the existence of an *evolutionarily stable population* because all types earn the same type fitness (and the definition of evolutionary stability requires residents to earn a strictly greater type fitness than mutants). In all the other cases, X_σ is a singleton and there does not exist any heterogeneous *evolutionarily stable population* with the residents playing diverse strategies. This observation is not a particular feature of a population of *homo oeconomicus* and *homo kantiensis* involved in a prisoners' dilemma. It remains valid for any residents playing a 2×2 symmetric game under uniformly-constant assortment. Recall that π_{ij} denotes the payoff when strategy x_i is played against strategy x_j. For $(x_1,x_2)\in B^{NE}(\theta_1,\theta_2,\lambda^\circ)$, we call $Q_{\pi_{1,2}} \equiv \pi_{11}-\pi_{21}-\sigma(\pi_{22}-\pi_{21})$ and $S_{\pi_{1,2}} \equiv \pi_{11}+\pi_{22}-\pi_{12}-\pi_{21}$. We have the following theorem:

Theorem 2 (Evolutionarily stable population)**.** *In a symmetric* 2×2 *fitness game under uniformly-constant assortment, let* $s^\circ = (\theta_1,\theta_2,\lambda^\circ)$ *be a heterogeneous population with* $\lambda^\circ \in (0,1)$.
The population s° *is evolutionarily stable against all types* $\theta_\tau \notin \Theta_{12}$ *if:*

- *When* $\sigma = 0$*: for all* $(x_1,x_2)\in B^{NE}(s^\circ)$, $x_1 = x_2 \in X_\sigma$ *and* $\beta_\sigma(x_1)$ *is a singleton.*
- *When* $\sigma > 0$*: for all* $(x_1,x_2)\in B^{NE}(s^\circ)$, $(x_1,x_2)\in X_\sigma^2$, $\beta_\sigma(x_1)$ *and* $\beta_\sigma(x_2)$ *are singleton, and for all* $(x_1,x_2)\in B^{NE}(s^\circ)$ *such that* $x_1 \neq x_2$, $Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}}) = \lambda^\circ$.

Conversely, if $(x_1,x_2)\in B^{NE}(s^\circ)$ *is a singleton such that* $(x_1,x_2)\notin X_\sigma^2$ *and if* Θ *is rich, then the population is not evolutionarily stable.*

Proof. In Appendix B.9. □

Theorem 2 tells us that if a heterogeneous population is evolutionarily stable, then the residents must play *Hamiltonian strategies* under uniformly-constant assortment. We will now study the case of state-dependent assortment and we will see that this property does not hold in the general case.

Evolutionary stability under state-dependent assortment

Under state-dependent assortment when *homo oeconomicus* and *homo kantiensis* earn the same type fitness in the state $s^\circ = (\theta_1,\theta_2,\lambda^\circ)$, the difference in type fitness between the residents and any mutant at the limit when the share of the mutant goes to zero is (Lemma 6):

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} = {} & [\alpha_\tau(1-\alpha_\tau)\sigma + (1-\alpha_\tau)\lambda^\circ(\phi_{12}-\sigma) + (1-\alpha_\tau)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot S_\pi \\ & + [(\alpha_\tau-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\pi^{CC}-\pi^{CD})\end{aligned}$$

We can first observe that if $\sigma = 1$ and if the mutants cooperate, they will always earn a greater type fitness than the residents. Indeed, in this setting, the mutants are matched between themselves earning $\Pi_{\theta_\tau} = \pi^{CC}$. On the other hand, the residents earn $\Pi_\theta < \pi^{CC}$. To see that, look at the type fitness of *homo kantiensis*: *homo kantiensis* individuals earn π^{CC} when matched with another *homo kantiensis* but they earn $\pi^{CD} < \pi^{CC}$ when matched with a *homo oeconomicus*. Consequently, there is a maximum value of assortativity allowing for a heterogeneous *evolutionary stable population*:

Proposition 3 (Evolutionary stability under state-dependent assortment)**.** *In a prisoners' dilemma, if Θ is rich then there exists $\bar{\sigma} < 1$ such that there does not exist a heterogeneous evolutionary stable population of homo oeconomicus and homo kantiensis for all $\sigma > \bar{\sigma}$.*

Proof. In Appendix B.10. □

Second, let $H : [0,1] \to \mathbb{R}$ be the function that maps the strategy played by the mutant α_τ to the difference in type fitness between the residents and any mutant at the limit, i.e. $H(\alpha_\tau) = \Pi_\theta - \Pi_{\theta_\tau}$. H is a polynomial of degree two. When H is concave, the function is strictly positive for all $\alpha_\tau \in [0,1]$ if and only if $H(0) > 0$ and $H(1) > 0$. In other words, when H is concave, it is sufficient to study what happens when the mutants defect and cooperate to know the sign of the difference in type fitness between the residents and any mutant at the limit. Thus, we have the following Theorem:

Theorem 3 (Evolutionarily stable population under state-dependent assortment)**.** *Let a population of homo oeconomicus (θ_1) and homo kantiensis (θ_2) in the state $s = (\theta_1, \theta_2, \lambda^\circ)$ involved in a prisoners' dilemma such that the type-fitness equality is satisfied.*
If $(\phi_{12} - \sigma) \notin [\Gamma\lambda^\circ, \Gamma(\lambda^\circ - 1)]$ and if Θ is rich, then the population is not evolutionarily stable. Conversely, when $S_\pi \geq 0$, the population is evolutionarily stable if $(\phi_{12} - \sigma) \in (\Gamma\lambda^\circ, \Gamma(\lambda^\circ - 1))$.

Proof. In Appendix B.11. □

Theorem 3 has a few implications. First, by contrast with the case of uniformly-constant assortment, we did not need the assumption that the mutants are not behaviorally-alike. Hence, thanks to state-dependent assortment, it is possible to find a heterogeneous *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* such that the residents resist the invasion of mutants that play like them.

Second, the Theorem implies that there does not exist a heterogeneous *evolutionarily stable population* when $\Gamma > 0$. Moreover, when $\Gamma = 0$, we need $\phi_{12} = \sigma$, and this case is analogous to uniformly-constant assortment already characterized in Theorem 1. The matching speed Γ governs which residents the mutants are more likely to meet. When $\Gamma > 0$, the mutants are additionally matched with *homo kantiensis* individuals, at the expense of encountering *homo oeconomicus* individuals. For instance, when $\Gamma = (1-\sigma)/\lambda^\circ$ and $\sigma = 0$, mutants always meet *homo kantiensis* individuals. Thus, the type-fitness of any mutant is a least $\Pi_{\theta_\tau} = \pi^{CC} > \Pi_\theta$

and the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable. Consequently, the matching speed Γ plays a central role in the analysis of evolutionary stability.

Finally, combined with Proposition 2 on type-fitness equality, Theorem 3 enables to know if a heterogeneous population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable in prisoners' dilemmas when $S_\pi \geq 0$. The Theorem remains valid in most cases when $S_\pi < 0$, and in particular when the minimum of H is reached in $\underline{\alpha_\tau} \notin (0,1)$. When $\underline{\alpha_\tau} \in (0,1)$, the condition for evolutionary stability is $H(\underline{\alpha_\tau}) > 0$, i.e. H has no real roots.

To illustrate Proposition 3 and Theorem 3, we focus on the same cases studied in Section 2.3.1:

1. We suppose that ϕ_{12} is linear: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$.
2. We suppose that ϕ_{12} is a U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$.
3. We suppose that ϕ_{12} is an inverse U-shaped parabola: for all $\lambda \in [0,1]$, $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$.

Moreover, we assume that $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$. This shape allows $p_{1\tau}$ and $p_{2\tau}$ to belong to $[0,1]$. Moreover, it means that $p_{1\tau} = 1 - \sigma$ and $p_{2\tau} = 0$, i.e. a mutant either meets a *homo oeconomicus* or another mutant, which increases the likelihood that the population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable.

For each case, we consider the same examples studied above and defined in Section 2.2.6: $\pi^{CD} = 0$, $\pi^{DD} = 1$, $\pi^{CC} = 4$, and (a) $\pi^{DC} = 6$, (b) $\pi^{DC} = 5$, (c) $\pi^{DC} = 4.5$.

1. When $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$:

   (a) $S_\pi < 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 1/6$. Moreover, for $\sigma < 0.4$, the residents earn a strictly greater type fitness than any mutant at the limit, and thus following the same arguments as in Theorem 1, the population is evolutionarily stable (see Figure 2.8a).

   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 7/24$. Moreover, for $\sigma < 0.46875$, the residents earn a strictly greater type fitness than any mutant (see Figure 2.8b). Hence, there exists a heterogeneous *evolutionarily stable population* when $\sigma < 0.46875$.

   (c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^\circ = 0.5$. They also earn a strictly greater type fitness than any mutant when $\sigma < 0.6$ and thus the population is evolutionarily stable (see Figure 2.8c).

2. When $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$:

   (a) $S_\pi < 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ -} \approx 0.130$ and $\lambda^{\circ +} \approx 0.943$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.368$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.966$ for $\lambda^{\circ +}$ (see Figure 2.9a).

   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness when $\lambda^{\circ -} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ +} = 0.5 + 0.25\sqrt{2}$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.360$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.890$ for $\lambda^{\circ +}$ (see Figure 2.9b).

(c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ -} \approx 0.157$ and $\lambda^{\circ +} \approx 790$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.355$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.825$ for $\lambda^{\circ +}$ (see Figure 2.9c).

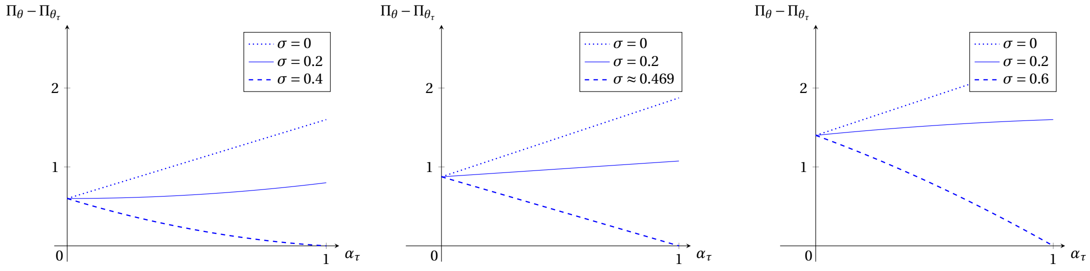
3. When $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$:

   (a) $S_\pi < 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ -} \approx 0.169$ and $\lambda^{\circ +} \approx 0.756$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.402$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.846$ for $\lambda^{\circ +}$ (see Figure 2.10a).

   (b) $S_\pi = 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ -} = 0.5 - 0.25\sqrt{2}$ and $\lambda^{\circ +} = 0.5 + 0.25\sqrt{2}$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.360$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.890$ for $\lambda^{\circ +}$ (see Figure 2.10b).

   (c) $S_\pi > 0$: *homo oeconomicus* and *homo kantiensis* earn the same type fitness for $\lambda^{\circ -} \approx 0.137$ and $\lambda^{\circ +} \approx 0.917$. The maximum assortativity allowing for a heterogeneous *evolutionarily stable population* is then $\bar{\sigma} \approx 0.342$ for $\lambda^{\circ -}$ and $\bar{\sigma} \approx 0.929$ for $\lambda^{\circ +}$ (see Figure 2.10c).

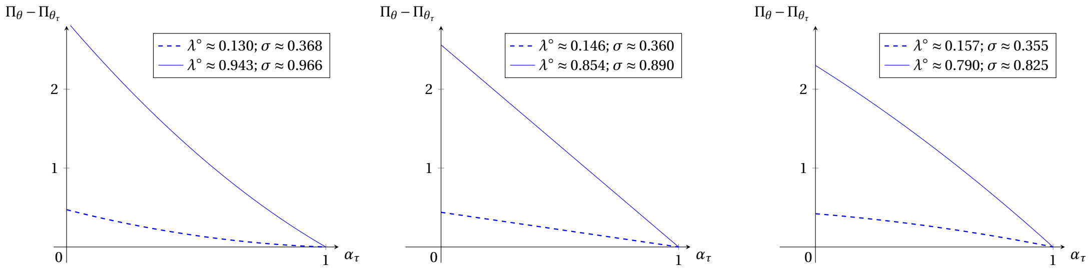
With our assumptions we find *evolutionarily stable populations* of *homo oeconomicus* and *homo kantiensis* in all games. This contrasts with the case of a uniformly-constant assortment, in which there is no *evolutionarily stable population* when S_π is negative. As discussed above, the heterogeneous *evolutionarily stable population* can also resist to the invasion of mutants that cooperate or defect.



(a) $S_\pi < 0$, $\lambda^\circ = 1/6$ **(b)** $S_\pi = 0$, $\lambda^\circ = 7/24$ **(c)** $S_\pi > 0$, $\lambda^\circ = 0.5$

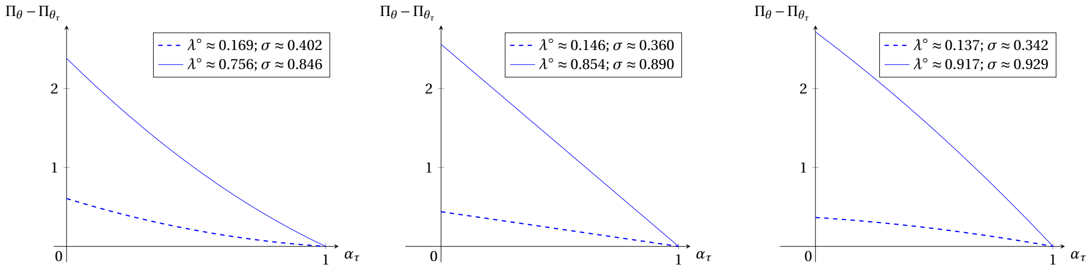
Figure 2.8 – Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 0.32 - 0.24\lambda$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.

Both *homo oeconomicus* and *homo kantiensis* are important for the evolutionary success of the population, but in a different way. On the one hand, *homo kantiensis* individuals drive up the average fitness of the population since Π_θ increases with the share of *homo kantiensis*. As a result, there exists a heterogeneous *evolutionarily stable population* for higher values of



(a) $S_\pi < 0$, **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 2.9 – Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.



(a) $S_\pi < 0$ **(b)** $S_\pi = 0$ **(c)** $S_\pi > 0$

Figure 2.10 – Difference in type fitness between a resident population of *homo kantiensis* and *homo oeconomicus* and the mutants in prisoners' dilemma, when the mutant share goes to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$ and $\Gamma = -(1-\sigma)/(1-\lambda^\circ)$.

assortativity (see Figures 2.9 and 2.10). On the other hand, *homo oeconomicus* individuals help to resist the invasion of mutants. Indeed, as discussed above, the population would not be evolutionarily stable when $\Gamma > 0$, i.e. when the mutants are additionally matched with *homo kantiensis* instead of *homo oeconomicus*.

2.3.3 On the robustness of *evolutionarily stable populations*

In the previous section, we have found an *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* in all games and all type-fitness equilibria under state-dependent assortment. By Definition 6, in an *evolutionarily stable population*, the two residents earn the same type fitness when there is no mutant in the population and they earn a strictly greater type fitness than a small share of mutants. However, the definition does not impose that the two residents earn the same type fitness when the mutants enter the population. Thus, the mutants could destabilize the residents. In this section, we discuss what happens then: could one type overcome the other because of the mutant entry?

To define the notion of evolutionary stability, we motivated the two conditions by referring to the framework of evolutionary game dynamics. In a dynamic game, the evolution of preferences depends on the difference between the fitness obtained and the average fitness in the population. Hence, when the two resident types get the same fitness in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$, their population share is stable (first condition of evolutionary stability). Moreover, in an *evolutionarily stable population*, the mutants will disappear first from the population since their type-fitness is strictly smaller than each of the residents in all Bayesian Nash equilibria in the states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ (second condition of evolutionary stability). Then, if the fitness of type θ_2 is greater than the fitness of θ_1, the population share of θ_2, i.e. λ, will increase. It will converge to λ° if $\lambda < \lambda^\circ$ and it will diverge if $\lambda > \lambda^\circ$. Reciprocally, if the fitness of type θ_2 is lower than the fitness of θ_1, the population share λ will decrease towards λ° if $\lambda > \lambda^\circ$. Consequently, to analyze if the mutant entry destabilizes the residents, we can simply look at how the difference in type fitness between the residents evolves around a small neighborhood of λ°.

Assuming that the assortment function ϕ_{12} is differentiable in λ°, we can define a robustness criterion:

Definition 9 (Robust equilibrium)**.** Let (x_1, x_2) be a BNE in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ such that θ_1 and θ_2 earn the same type fitness. The equilibrium is called robust if:

$$\left.\frac{\partial\left(\Pi_{\theta_1} - \Pi_{\theta_2}\right)}{\partial\lambda}\right|_{\lambda=\lambda^\circ} > 0$$

By extension, a population will be called robust in the state s° if all BNE in s° are robust.

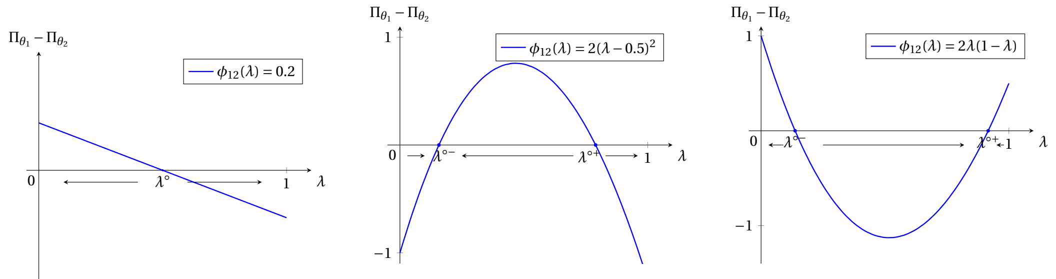
When the condition of Definition 9 is satisfied, the type fitness of θ_2 becomes smaller than the type fitness of θ_1 when the share of θ_2 increases. Similarly, the type fitness of θ_2 becomes larger than the type fitness of θ_1 when the share of θ_2 decreases. Hence, in a dynamic setting when λ is close to λ°, the population would converge towards λ°.

We illustrate the definition looking back at our previous examples, for $S_\pi > 0$:

(a) Under uniformly-constant assortment $\phi_{12}(\lambda) = \sigma = 0.2$, for $\lambda^\circ = 0.5$ the population satisfies the type-fitness equality. However, $\partial\left(\Pi_{\theta_1} - \Pi_{\theta_2}\right)/\partial\lambda = -(1-\sigma)S_\pi < 0$ so that any deviations in λ would destabilize the population (see Figure 2.11a). If $\lambda < \lambda^\circ$, the population evolves towards a homogeneous population of *homo oeconomicus.* On the contrary, if $\lambda > \lambda^\circ$, the population evolves towards a homogeneous population of *homo kantiensis.*

(b) Under state-dependent assortment $\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$, there are two population shares $\lambda^{\circ-}$ and $\lambda^{\circ+}$ satisfying type-fitness equality. However, only the low-cooperation share $\lambda^{\circ-}$ is robust (see Figure 2.11b). When $\lambda \in [0, \lambda^{\circ+})$, the population evolves towards the heterogeneous population in the state $s^{\circ-} = (\theta_1, \theta_2, \lambda^{\circ-})$. When $\lambda > \lambda^{\circ+}$, the population evolves towards a homogeneous population of *homo kantiensis.*

(c) Under state-dependent assortment $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$, there are two population shares $\lambda^{\circ-}$ and $\lambda^{\circ+}$ satisfying type-fitness equality. However, only the high-cooperation share $\lambda^{\circ+}$ is robust (see Figure 2.11c). When $\lambda < \lambda^{\circ-}$, the population evolves towards a homogeneous population of *homo oeconomicus*. When $\lambda \in (\lambda^{\circ-}, 1]$, the population evolves towards the heterogeneous population in the state $s^{\circ+} = (\theta_1, \theta_2, \lambda^{\circ+})$.



(a) Non-robust

(b) $\lambda^{\circ-}$ Robust
$\lambda^{\circ+}$ Non-robust

(c) $\lambda^{\circ-}$ Non-robust
$\lambda^{\circ+}$ Robust

Figure 2.11 – Robustness in prisoner's dilemma ($S_\pi > 0$) between *homo oeconomicus* (Π_{θ_1}) and *homo kantiensis* (Π_{θ_2})

Note that under uniformly-constant assortment, there does not exist any heterogeneous *evolutionarily stable population* when S_π is negative (Theorem 1). Furthermore, as illustrated in Figure 2.11a, the population is not robust when S_π is positive. Hence, there does not exist any robust heterogeneous *evolutionarily stable population* under uniformly-constant assortment:

Proposition 4 (Robust evolutionarily stable population)**.** *In a prisoners' dilemma under uniformly-constant assortment, there does not exist any robust heterogeneous evolutionarily stable population of homo oeconomicus and homo kantiensis.*

Proof. In Appendix B.12. □

This result highlights once again the importance of assortment in the analysis of evolutionary stability. Under state-dependent assortment, we found for all our examples one state such that the population is evolutionarily stable and robust. When the assortment ϕ_{12} is a U-shaped parabola ($\phi_{12}(\lambda) = 2(\lambda - 0.5)^2$), only the low-cooperation state is robust. When the assortment ϕ_{12} is an inverse U-shaped parabola ($\phi_{12}(\lambda) = 2\lambda(1-\lambda)$), only the high-cooperation state is robust. Thus, the robustness criterion allows to select which equilibria are more likely. These results are summarized in the following tables.

Table 2.2 – Robust *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* involved in a prisoner's dilemma $S_\pi < 0$

Equilibrium λ°	0.5	1/6	0.130	0.943	0.169	0.756
$\phi_{12}(\lambda)$	1/3	$0.32 - 0.24\lambda$	$2(\lambda - 0.5)^2$		$2\lambda(1-\lambda)$	
σ	1/3	0-0.4	0-0.368	0-0.966	0-0.402	0-0.846
Γ	0	$-(1-\sigma)/(1-\lambda)$				
Evolutionarily stable	No	Yes	Yes	Yes	Yes	Yes
Robust	Yes	Yes	Yes	No	No	Yes

Table 2.3 – Robust *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* involved in a prisoner's dilemma $S_\pi = 0$

Equilibrium λ°	0-1	7/24	0.146	0.854	0.146	0.854
$\phi_{12}(\lambda)$	0.25	$0.32-0.24\lambda$	$2(\lambda-0.5)^2$		$2\lambda(1-\lambda)$	
σ	0.25	0-0.468	0-0.360	0-0.890	0-0.360	0-0.890
Γ	0	$-(1-\sigma)/(1-\lambda)$				
Evolutionarily stable	No	Yes	Yes	Yes	Yes	Yes
Robust	No	Yes	Yes	No	No	Yes

Table 2.4 – Robust *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* involved in a prisoner's dilemma $S_\pi > 0$

Equilibrium λ°	0.5	0.5	0.157	0.790	0.137	0.917
$\phi_{12}(\lambda)$	0.2	$0.32-0.24\lambda$	$2(\lambda-0.5)^2$		$2\lambda(1-\lambda)$	
σ	0.2	0-0.6	0-0.355	0-0.825	0-0.342	0-0.929
Γ	0	$-(1-\sigma)/(1-\lambda)$				
Evolutionarily stable	Yes	Yes	Yes	Yes	Yes	Yes
Robust	No	Yes	Yes	No	No	Yes

2.4 Homogeneous vs heterogeneous *evolutionarily stable populations*

In this chapter, we have expanded the classical framework of evolutionary stability to analyze whether a heterogeneous population of *homo oeconomicus* and *homo kantiensis* could be favored by evolution. In this section, we discuss some implications of introducing heterogeneity in the resident population, by reviewing the differences between a homogeneous and a heterogeneous *evolutionarily stable population.*

2.4.1 Favored preferences and strategies

Adapting the framework of evolutionary stability formally established by Maynard Smith and Price (1973) for strategies, Alger and Weibull (2013) proved the evolutionary stability of a particular type of preference, *homo hamiltonensis* in a homogeneous population. As first expectation, we could have hypothesized that a heterogeneous *evolutionarily stable population* would "on average" have a *homo hamiltonensis* preference. In other words, an intuitively good candidate for a heterogeneous *evolutionarily stable population* would be a population composed by fully-selfish and fully-moral individuals with a share σ of fully moral individuals in order to "mimic" a *homo hamiltonensis* utility. However, such a population is not evolutionarily stable in most cases.[21]

Instead, Theorem 2 shows that a heterogeneous *evolutionarily stable population* under

[21] The only case in which this population is evolutionarily stable is when $\sigma = \lambda$ and σ is a solution of $\sigma = (\pi^{DD} - \pi^{CD} - \sigma(\pi^{CC} - \pi^{DC}))/((1-\sigma)S_\pi)$.

uniformly-constant assortment depends on *Hamiltonian strategies*. In other words, evolution favors the same strategies in a homogeneous population and in a heterogeneous population under uniformly-constant assortment. However, the introduction of state-dependent assortment allows for the existence of heterogeneous *evolutionarily stable population* in all games studied, so that the residents do not need to play *Hamiltonian strategies* anymore. Hence, the heterogeneous framework favors a greater diversity of preferences and strategies.

2.4.2 Equilibrium implications

In the classical setting of a homogeneous population, all resident individuals play the same strategy. We show that this characteristic is not necessary for evolutionary stability by proving the existence of a heterogeneous population exhibiting diverse strategies played by resident individuals without infringing the evolutionary stability. For example, in the prisoner's dilemma when $S_\pi > 0$, all *homo hamiltonensis* individuals either cooperate or defect, i.e. they all behave as a *homo oeconomicus* and defect, or they all behave as a *homo kantiensis* and cooperate. Yet, Theorem 1 establishes the existence of a heterogeneous *evolutionarily stable population* with a share of defectors *homo oeconomicus* and of cooperators *homo kantiensis*.

This last result is more in line with empirical observations. In single trial public goods experiments for instance, results display a 40% to 60% contribution to the public good (Marwell and Ames, 1981; Dawes and Thaler, 1988). A population of *homo hamiltonensis* all playing a mixed strategy in a prisoner's dilemma could support this empirical observation when $S_\pi < 0$ but not when $S_\pi > 0$ (Lemma 8). In the latter case, only a heterogeneous population would justify the observations.

2.4.3 Assortative matching and Nash equilibrium

The introduction of assortative matching between preferences has a key implication when studying and interpreting equilibria in games. In his thesis, John Nash discussed two interpretations of a mixed Nash equilibrium (Nash, 1950, 1951). In the first interpretation, an individual randomizes his play before acting, for instance by throwing a dice or a coin. In the second, called "mass-action", individuals of a large population play one of the pure strategies composing the mixed equilibrium with the share of people playing each strategy being equal to the weight of the strategy in the equilibrium.[22] Similarly, in the original and static evolutionary game theory framework (Maynard Smith, 1974), a mixed *evolutionarily stable strategy* can either describe a "monomorphic" population of identical individuals randomizing their behavior, or a heterogeneous population (called "polymorphic" in biology) of several types of individuals, each type playing a pure strategy. Under uniform random matching, the two interpretations are equivalent. Thus, the static framework could not distinguish between a monomorphic and a polymorphic population, which led to the emergence of the evolutionary game dynamics framework (Bergstrom and Godfrey-Smith, 1998). However, when

[22] See also Leonard (1994) and Weibull (1994) for a discussion of the mass-action interpretation of Nash equilibria.

the matching is assortative, a monomorphic and a polymorphic population would not yield the same equilibrium, as already observed by Grafen (1979) and Hines and Maynard Smith (1979). In other words, the first and second interpretation of a mixed equilibrium are no longer equivalent when a distinct preference is associated to each strategy.

2.4.4 Context-based preferences

A key property in the case of a homogeneous population is the evolutionary stability of the *homo hamiltonensis* preference regardless of the game being played. In other words, as long as the assortativity is set and constant, in any game between assortatively matched individuals, only those behaviorally alike to *homo hamiltonensis* will resist mutant invasion. This property does not hold anymore in a heterogeneous population. Indeed, we have shown that the evolutionary stability depends on the game being played. Specifically, we find that both the assortment properties and the game payoffs determine whether a heterogeneous population is evolutionarily stable. For instance, in a prisoner's dilemma under uniformly-constant assortment, the evolutionary stability of a population of *homo oeconomicus* and *homo kantiensis* individuals depends on the sign of S_π and the value of assortativity σ (Theorem 1).

Hence, the prevailing preferences in a population depend on the context. This finding is in line with earlier research stating that the economic environment determines the prevalence of self-interested or altruistic behaviors (Bester and Güth, 1998) and of self-interested or fair behaviors (Fehr and Schmidt, 1999). Similarly, times preferences and attitudes toward risk are shaped by the environment (Netzer, 2009). Empirical evidence also suggests that choices and preferences can change according to the context (Tversky and Simonson, 1993; Rieskamp et al., 2006; Masatlioglu et al., 2012; Bordalo et al., 2013). As examples, economic crises modify the attitude toward risk (Schildberg-Hörisch, 2018) and the social, economic and institutional settings affect cooperative behaviors (Shogren and Taylor, 2008). In our framework, a socio-economic shock would translate into a change in the payoffs and in the homophily (i.e. the assortment), which would, in turn, affect the prevailing preferences in the population.

This dependence on the context has significant implications for empirical testing. Since the game and the context affect the behavior of agents, experiments should give particular attention to the conditions under which experiments are performed (statement of payoffs, cost of actions, available options, ties between subjects, etc.). While empirical behavioral research often aims at finding the parameters of the preferences of individuals, it would be an interesting challenge to try to estimate how diverse a population is. Considering a distribution of *homo moralis* with different morality coefficients, what is the shape of this distribution? The framework developed in this paper could also be tested in lab experiments. For instance, in the case of the prisoner's dilemma, does our simplified model explain the share of individuals cooperating? Is there assortment between individuals with similar preferences, and if so, what is the shape of assortment functions in different contexts and cultures? In all these experiments, the choice of payoffs in the game is central, since different payoffs lead to different evolutionary stability profiles.

2.4.5 Assortativity and evolutionary stability

Even though *homo hamiltonensis* is the favored preference in a homogeneous population regardless of the game being played, this result requires a constant assortativity σ. However, if the assortativity σ evolves, then all resident individuals become vulnerable to the entry of mutants. For instance, in the prisoners' dilemma previously studied with $S_\pi < 0$ and $\sigma = 0.25$, the homogeneous *evolutionarily stable population* consists of *homo moralis* with a morality coefficient $\kappa = 0.25$ and the only *Hamiltonian strategy* is the mixed strategy $x_\sigma = (S_\pi - R_\pi)/((1+\sigma)S_\pi) = 0.8$ (by Lemma 8). Now if the assortativity slightly decreases such that $\sigma = 0.24$, the only evolutionarily stable strategy is the mixed strategy $x_\sigma \approx 0.834$ and *homo moralis* with morality $\kappa = 0.25$ is no longer evolutionarily stable.

By contrast, in a heterogeneous *evolutionarily stable population*, types who would not be evolutionarily stable alone mutually contribute to resist the invasion of mutants, and they are less sensitive to variations in the assortativity. For instance in the same game $S_\pi < 0$, *homo oeconomicus* and *homo kantiensis* would not be evolutionarily stable in a homogeneous population. But when the assortment is state-dependent with $\phi_{12}(\lambda) = 2\lambda(1-\lambda)$ and $\Gamma = -(1-\sigma)/(1-\lambda)$, they can be part of a heterogeneous *evolutionarily stable population* and for $\lambda^\circ = 0.756$, they can resist any mutant invasion when the assortativity is smaller than $\bar{\sigma} = 0.846$ (see Figure 2.10c). Moreover, this heterogeneous *evolutionarily stable population* also resists the invasion of behaviorally-alike mutants, i.e. mutants playing like the residents. This is not the case in a homogeneous population. Consequently, our findings provide theoretical evidences in favor of the observed heterogeneity of preferences.

2.5 Toward a greater diversity of preferences

Throughout this chapter, we analyzed the evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis* involved in a prisoners' dilemma. The choice of *homo oeconomicus* and *homo kantiensis* was made for several reasons. First, since a particular type of *homo moralis* is favored by evolution in a homogeneous population, *homo oeconomicus* and *homo kantiensis* were good candidates to be part of a heterogeneous *evolutionarily stable population*. Second, *homo oeconomicus* and *homo kantiensis* individuals are committed to a strategy, i.e. they always play defect and cooperate no matter their population share. This property allows us to simplify the analysis. However, the framework developed in this chapter is more general and could apply to any types. In this section, we go beyond a population of *homo oeconomicus* and *homo kantiensis* discussing how the introduction of heterogeneity allows for a greater diversity of preferences.

2.5.1 Mixed strategies and evolutionary stability

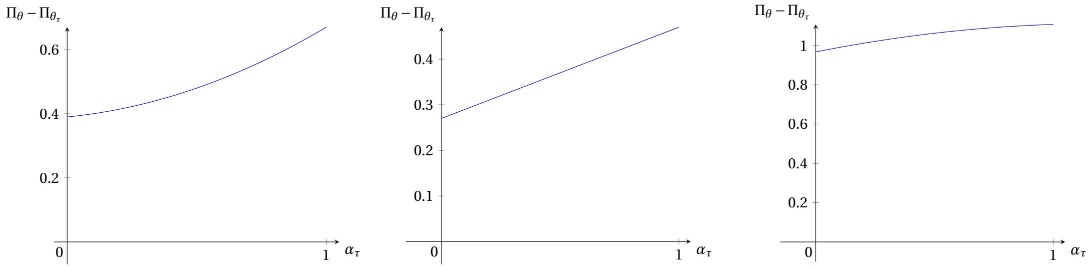
Although we focused on a population of *homo oeconomicus* and *homo kantiensis* playing the pure strategies "defect" and "cooperate", the introduction of state-dependent assortment enables the existence of heterogeneous *evolutionarily stable populations* in which the residents play mixed strategies. In other words, the residents in a heterererogeneous *evolutionarily stable population* are not necessarily behaviorally-alike to *homo kantiensis* or *homo oeconomicus.*

For simplicity, we will assume that the types are committed to their strategy. For all $\lambda \in [0,1]$, let $\phi_{12}(\lambda) = 0.24\lambda + 0.08$, $\Gamma = (1-\sigma)/\lambda$ and $\sigma = 0.2$. Going back to our three examples, we can find heterogeneous *evolutionarily stable populations* in which the residents play mixed strategies in all games previously studied:

(a) When $S_\pi < 0$: Suppose that individuals θ_1 cooperate with probability $\alpha_1 = 0.78$ and that individuals θ_2 cooperate with probability $\alpha_2 = 0.1$. For $\lambda^\circ = 5/6$, $\phi_{12}(\lambda^\circ) = 0.28$, and θ_1 and θ_2 get the same type fitness $\Pi_\theta \approx 1.79$. The residents also earn a greater type fitness than any mutant at the limit and the population is therefore evolutionarily stable (see Figure 2.12a).

(b) When $S_\pi = 0$: Suppose that individuals θ_1 cooperate with probability $\alpha_1 = 0.8$ and individuals θ_2 cooperate with probability $\alpha_2 = 0.4$. For $\lambda^\circ = 17/24$, $\phi_{12}(\lambda^\circ) = 0.25$, and θ_1 and θ_2 get the same type fitness $\Pi_\theta = 2.55$. The residents also earn a greater type fitness than any mutant at the limit rendering the population evolutionarily stable (see Figure 2.12b).

(c) When $S_\pi > 0$: Suppose that individuals θ_1 cooperate with probability $\alpha_1 = 0.85$ and individuals θ_2 cooperate with probability $\alpha_2 = 0.15$. For $\lambda^\circ = 0.5$, $\phi_{12}(\lambda^\circ) = 0.2$, individuals θ_1 and θ_2 get the same type fitness $\Pi_\theta \approx 2.387$.The residents also earn a greater type fitness than any mutant at the limit and the population is evolutionarily stable (see Figure 2.12c).



(a) $S_\pi < 0$, $\lambda^\circ = 5/6$

(b) $S_\pi = 0$, $\lambda^\circ = 17/24$

(c) $S_\pi > 0$, $\lambda^\circ = 0.5$

Figure 2.12 – Type-fitness difference between a resident population playing mixed strategies ((a) $(\alpha_1, \alpha_2) = (0.78, 0.1)$; (b) $(\alpha_1, \alpha_2) = (0.8, 0.4)$; (c) $(\alpha_1, \alpha_2) = (0.85, 0.15)$) and the mutants involved in a prisoners' dilemma, when the mutant share tends to zero, depending on the strategy played by mutants (α_τ), under state-dependent assortment: for all $\lambda \in [0,1]$ $\phi_{12}(\lambda) = 0.24\lambda + 0.08$, $\Gamma = (1-\sigma)/\lambda$ and $\sigma = 0.2$

These examples illustrate the variety of possible heterogeneous *evolutionarily stable populations* under state-dependent assortment. They also confirm that resident individuals can play strategies outside the set of *Hamiltonian strategies* and that mixed strategies can be observed in a heterogeneous *evolutionarily stable population.* Nevertheless, it is worth mentioning that the shapes of ϕ_{12} and Γ were set arbitrarily. A more in-depth analysis of state-dependent assortment is needed to derive more generic results and to better understand the conditions under which heterogeneous *evolutionarily stable populations* exist in this case.

2.5.2 Assortativity dependent on mutants

As discussed in Section 2.4.5, the evolutionary stability of *homo hamiltonensis* in a homogeneous population requires a constant assortativity σ. Another assumption needed is that the assortativity is exogenous. In other words, σ is the same for all the mutants. However, since the assortativity could have both genetic and cultural determinants, it is likely that it depends on the mutant type. In this case, there does not exist any homogeneous *evolutionarily stable population*, as proved by Newton (2017). In particular, the homogeneous population could be invaded by *homo oeconomicus* with assortativity $\sigma_0 = 0$ or by *homo kantiensis* with assortativity $\sigma_1 = 1$.

Even though heterogeneous *evolutionarily stable populations* are more resistant to a change in assortativity, they are not immune to mutant-dependent assortativity. Indeed, as shown in Proposition 3 there is a maximum value of σ allowing for the existence of a heterogeneous *evolutionarily stable population*. Hence, when the assortativity is mutant-dependent, the resident population is always vulnerable to mutant invasion. As a result, the population would be even more diverse and in perpetual evolution.

2.5.3 Unobserved diversity of preferences: on altruism, empathy and imitation

In Theorems 1 and 3, we have detailed the conditions under which a population of selfish *homo oeconomicus* and fully-moral *homo kantiensis* can be evolutionarily stable in a prisoner's dilemma under uniformly-constant assortment. This result can be extended to the behaviorally-alike of *homo oeconomicus* and *homo kantiensis*. In particular, individuals caring only for the payoff of others such as fully-altruistic or fully-empathetic individuals would always cooperate in a prisoner's dilemma.[23] Thus, they can be part of a heterogeneous *evolutionarily stable population* with *homo oeconomicus* individuals. In fact, in the case of a homogeneous population, Alger et al. (2018) have shown that the favored preference by evolution under weak selection consists of a selfish, a moral and an altruistic component.

Are individuals more driven by morality or altruism? Our framework provides a theoretical-justification for the observed diversity of behaviors and preferences but cannot answer this question. Thus, it would be interesting to empirically test which social preferences explain individuals' choices better. For instance, Miettinen et al. (2017) have recently shown that *homo moralis* has a higher explanatory power than altruistic preferences in a sequential prisoners' dilemma. However, scientists can only observe the strategies chosen by individuals and not

[23] The utility of fully-altruistic and fully-empathetic individuals is $u(x, y) = \pi(y, x)$. See also Alger and Weibull (2017) for a discussion of the strategic behaviors of moralists and altruists.

their true preferences. As discussed above, these strategies are context-dependent. Hence, further investigation varying the games and the context of the experiment would help identify individual preferences with greater precision and better understand the individual motives behind the observed decisions.

Furthermore, under uniformly-constant assortment, the population is not evolutionarily stable against behaviorally-alike mutants. Hence, when the mutants have information about the resident types, they have incentives to imitate the strategies played by the residents. This means that if we reverse the questions looking at the mutant's perspective instead of the resident's, evolution would favor imitative behaviors and preferences. This is in line with empirical evidence revealing imitative behaviors in a variety of contexts such as the contribution to public goods (Carpenter, 2004; Alpizar et al., 2008; Shang and Croson, 2009), markets (Cont and Bouchaud, 2000; Selten and Apesteguia, 2005), environmental and energy conservation (Goldstein et al., 2008; Allcott, 2011) or sustainable food consumption (Vermeir and Verbeke, 2006). The conformity to social and cultural norms has been observed not only in humans (Bovard Jr, 1953) but also chimpanzees (Whiten et al., 2005), and has inspired models on conformism (Akerlof, 1997) or on reciprocity (Fehr and Gächter, 2000). Consequently, even if our study focuses on a heterogeneous moral population, the picture depicted here is not complete and they are many other important drivers of individuals' behaviors.

2.6 Lessons learned

Individuals exhibit a wide heterogeneity in their social preferences (Falk et al., 2018). Following this empirical observation, we extended the classical framework of evolutionary stability of preferences by allowing heterogeneity in individual preferences in the context of assortative interactions with imperfect information. We generalized the concept of assortment function to define an assortment matrix modeling homophily between the different types of preferences in a population. We proved that there exists heterogeneous *evolutionarily stable population* composed of fully-selfish individuals, *homo oeconomicus*, and fully-moral ones, *homo kantiensis*, for some but but not all games and assortment structures. We showed that in the case of uniformly-constant assortment, individuals in a heterogeneous *evolutionarily stable population* should play *Hamiltonian strategies*, the strategies played by the evolutionarily stable *homo moralis* in a homogeneous population. By contrast, state-dependent assortment allows a greater diversity and enhances the robustness of *evolutionarily stable populations.* In a heterogeneous environment, individuals do not necessarily play the same strategy. Thus, our work helps in understanding the driving forces behind strategic behavior such as cooperation and defection in social dilemma or the diverse contributions to public goods. In the following Chapter, we will show how a heterogeneous moral population can shed light on why some individuals are willing to perform environmental-friendly actions.

The setting developed in this chapter provides a theoretical framework pushing the development of analyzes accounting for a diversity of preferences under assortative matching. Many extensions and improvements can be undertaken to deepen the understanding of heterogeneous populations. First, further exploring the case of state-dependent assortment, of which we analyzed three different cases, is key to better comprehend the role assortment plays in allowing for the diversity of preferences. The assortment could be rendered endogenous by including informational and strategic features into the game. It would be interesting to study how to define assortment in the case of a distribution of preferences in order to reconcile our framework with the one of Dekel et al. (2007). Moreover, the analysis of a heterogeneous *evolutionarily stable population* could be extended to finite games with more than two pure strategies and more than two resident types, and to infinite games. Would *Hamiltonian strategies* still be favored under uniformly-constant assortment? Finally, in our analysis, we favored a static framework because we investigated under which conditions a heterogeneous population is evolutionarily stable to the invasion of a mutant preference. It would be helpful to analyze how the preferences in a heterogeneous population evolve under assortative matching using an evolutionary game dynamics framework. We expect that some equilibria we found in the static case could not be reached in a dynamic setting depending on the evolutionary process.

3 Why do individuals care for Nature?

Disclaimer: This chapter draws from a working paper written with Charles Ayoubi. I would like to warmly thank Charles for his help in the elaboration of this chapter. I would also like to acknowledge Prof. Philippe Thalmann, Dr. Damien Ackerer and Sophia Ding for their useful feedback. This research has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 730459, EUCalc project.[1]

3.1 Motivation

Individuals often engage voluntarily in costly pro-environmental actions even though their efforts have little or even negligible impacts on the improvement of the environmental quality. However, the propensity to act in favor of the environment differs across individuals and countries. This has been observed in many contexts such as the willingness to pay a premium to purchase green electricity (Sundt and Rehdanz, 2015) and sustainable food (Moon et al., 2002). Households' recycling efforts also vary among individuals (Bruvoll et al., 2002) and countries, as illustrated in Figure 3.1. The objective of this Chapter is to better comprehend why some individuals care for Nature in order to help policy makers design effective environmental policies.

The *homo oeconomicus* type of preference fails to explain environmental-friendly behaviors. If individuals were selfish, they would prefer the cheapest and "dirty" option. More interestingly, when the action of each individual has a negligible impact on the environmental quality, altruistic motives cannot explain these behaviors either. For instance, the greenhouse gas emissions of each individual has an insignificant impact on climate change, so that individuals' efforts have a negligible effect on the well-being of others. By contrast, the *homo moralis* preference is a good candidate to understand pro-environmental actions since moral individuals consider what happens when everybody acts like them when making decisions. Furthermore, heterogeneous moral populations can be favored by evolution, as shown in Chapter 2.

[1] To learn more about EUCalc: http://www.european-calculator.eu/

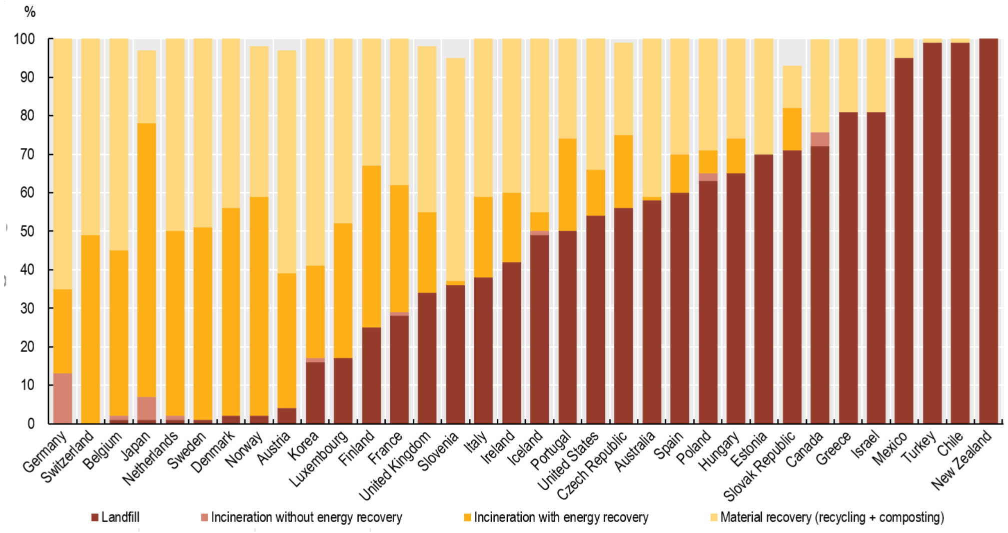


Figure 3.1 – Municipal waste disposal and recovery shares in OECD countries. Source: OECD (2015), Environment at a Glance 2015: OECD Indicators, OECD Publishing, Paris. Database: "Waste: Municipal waste", OECD Environment Statistics, http://dx.doi.org/10.1787/data-00601-en

Building on these observations, we design a simple framework with heterogeneous moral agents involved in a social dilemma, where each individual action has no effect on the total contribution.[2] Our model sheds light on the motives behind pro-environmental behaviors and on the determinants of the level of cooperation in a population. By analyzing the influence of individuals' beliefs on their decisions, we also show why financial incentives could fail in some cases and we offer policy recommendations to promote environmental-friendly behaviors.

Although there exist many empirical studies documenting individuals' willingness to engage in pro-environmental actions, there are few theoretical models explaining why they do so. Most of the literature relies on the concept of "warm-glow", which reflects the idea that individuals get a positive utility from doing an action that is perceived as "good" by society (Andreoni, 1989, 1990).[3] For instance, Abbott et al. (2013) examine recycling efforts by including the time spent recycling in the household's utility function. However, this modeling approach is *ad-hoc* and lacks theoretical foundations. Brekke et al. (2003) and Nyborg et al. (2006) incorporate self-image in individuals' utility, such that the morally superior (green) alternative yields a self-image benefit. Nevertheless, they assume that all individuals are identical.

The rest of the Chapter is organized as follows. In Section 3.2 we introduce the model. In Section 3.3, we analyze why some individuals cooperate in social dilemmas and we study what the level of cooperation in the population is. In Section 3.4, we investigate how individuals' perception alter their decisions. In Section 3.5, we provide some applications to the purchase of green electricity, of electric vehicles and of sustainable food to illustrate how our model can

[2] The literature often refers to the notion of "public goods" instead of "social dilemma". We favor here the term "social dilemma" to avoid the confusion with public goods experiments in which individuals' choices modify the total contribution.

[3] "Warm-glow" is frequently called "impure altruism".

be applied and extended in a variety of contexts. In Section 3.6, we examine the effectiveness of policies to foster pro-environmental behaviors. Finally, we recap our findings and we discuss the limitations of our model in Section 3.7.

3.2 Model

In this section, we present the model and the main definitions. We consider a large population consisting of a continuum of individuals $i \in I = [0,1]$ involved in a social dilemma (Section 3.2.1). Individuals have *homo moralis* preferences and the population is heterogeneous, i.e. individuals have different degrees of morality (Section 3.2.2).

3.2.1 Social dilemma

In a social dilemma, the society is better-off if all individuals cooperate (strategy C). However, each individual has an incentive to defect (strategy D). With only pure strategies, the strategy set is therefore $X = \{C,D\}$. For all individuals $i \in [0,1]$, we call $x_i \in [0,1]$ the level of cooperation of individual i. In others words, $x_i = 1$ if individual i cooperates and $x_i = 0$ if she defects. The strategy set can then alternatively be written as $X = \{0,1\}$. We can then define the (average) cooperation share in the population $\bar{x} = \int_{i\in I} x_i \, d\mu$, with μ a density for the population I. Thus, $\bar{x}$ represents the share of cooperators in the population.

Property 6 (Atomicity)**.** Since the population is large, the average level of cooperation in the population is unaffected by the action of a single individual:

$$\forall j \in I: \quad \bar{x} = \int_{i\in I} x_i \, d\mu = \bar{x}_{-j} = \int_{I-\{j\}} x_i \, d\mu$$

This property is called atomicity.

For any cooperation share $\bar{x} \in [0,1]$, an individual i playing $x_i \in \{0,1\}$ gets a material payoff $\pi_i(x_i, \bar{x})$, where we assume $\pi_i : \{0,1\} \times [0,1] \to \mathbb{R}$ to be continuous and differentiable in $\bar{x}$ for all individuals $i \in I$. In other words, the payoff of an individual i depends both on her strategy x_i and on the others' strategies through the share $\bar{x}$ of individuals cooperating in the population. Moreover, the setting of a social dilemma implies three main assumptions:

Assumption 1 (Cooperation share and individual payoff)**.** Individuals' payoffs are strictly increasing with the cooperation share:

$$\forall i \in I, \forall \bar{x} \in [0,1], \forall x_i \in \{0,1\}: \quad \partial \pi_i(x_i, \bar{x})/\partial \bar{x} > 0$$

Assumption 2 (Social benefit of cooperation)**.** Individuals get a strictly greater payoff if everybody cooperates than if everybody defects, i.e.:

$$\forall i \in I: \quad \pi_i(C,1) > \pi_i(D,0)$$

Assumption 3 (Individual incentive to defect)**.** For any value of the cooperation share, each individual is better-off by defecting instead of cooperating:

$$\forall i \in I, \forall \bar{x} \in [0,1]: \quad \pi_i(D,\bar{x}) > \pi_i(C,\bar{x})$$

The social dilemma defined above is illustrated in Figure 3.2.

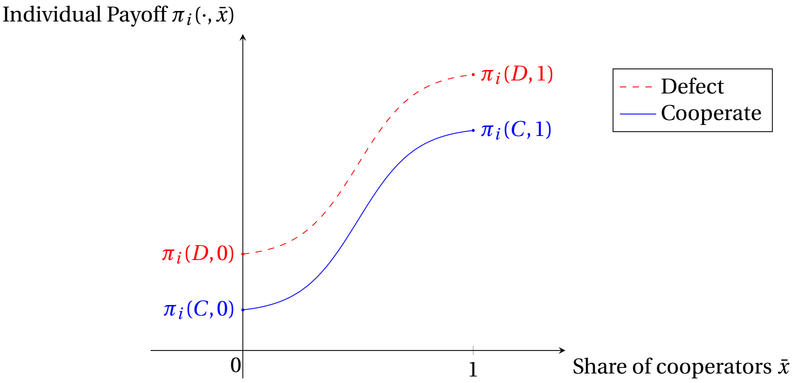


Figure 3.2 – Illustrative Social Dilemma

In the analysis (Section 3.3), we will see that two key variables drive the decisions of individuals.

Definition 10 (Individual cost)**.** For all $i \in I$, we call individual cost IC_i the difference in the material payoff between defecting and cooperating for a given cooperation share $\bar{x}$:

$$IC_i(\bar{x}) = \pi_i(D,\bar{x}) - \pi_i(C,\bar{x})$$

The individual cost IC_i is the individual cost of cooperating (or incentive to defect) of individual $i \in [0,1]$. From Assumption 3, we know that for all $i \in I$ and for any $\bar{x} \in [0,1]$, $IC_i(\bar{x}) > 0$.

Definition 11 (Social benefit)**.** For all $i \in I$, we call social benefit SB_i, the difference in the material payoff of individual i between a situation of full cooperation in the population and a situation of no cooperation:

$$SB_i = \pi_i(C,1) - \pi_i(D,0)$$

The social benefit of individual i SB_i is the difference between her payoff if she cooperates in a population of cooperators and her payoff if she defects in a population of defectors. From Assumption 2, we know that for all $i \in I$, $SB_i > 0$.

Finally, we introduce a particular case which offers an interesting framework for the analysis:

Definition 12 (Uniform Cost-Benefit)**.** We will say that the cost-benefit structure is uniform when the values of individual cost and social benefit are the same for all individuals in the population and independent of the average cooperation in the population $\bar{x}$:

$$\begin{cases} \forall i \in I, \bar{x} \in [0,1]: \quad IC_i(\bar{x}) = IC \\ \forall i \in I: \quad SB_i = SB \end{cases}$$

3.2.2 Population

Individuals are utility maximizers and they all have a *homo moralis* type of preference. We introduced *homo moralis* in Chapter 2: an individual i with a *homo moralis* preference maximizes a convex combination of her classical selfish payoff, with a weight $(1-\kappa_i)$, and of her "moral" payoff, defined as the payoff she would get if everybody played like her, with a weight κ_i. However, the definition of Section 2.2.5 was appropriate in the context of bilateral interactions such as prisoners' dilemma. Thus we need to adjust the definition for individuals involved in a social dilemma:

Definition 13 (Homo moralis)**.** An individual is a *homo moralis* if her utility function is of the form:

$$u_{\kappa_i}(x_i, \bar{x}) = (1-\kappa_i) \cdot \pi_i(x_i, \bar{x}) + \kappa_i \cdot \pi_i(x_i, x_i)$$

where $\kappa_i \in [0,1]$ is her degree of morality.

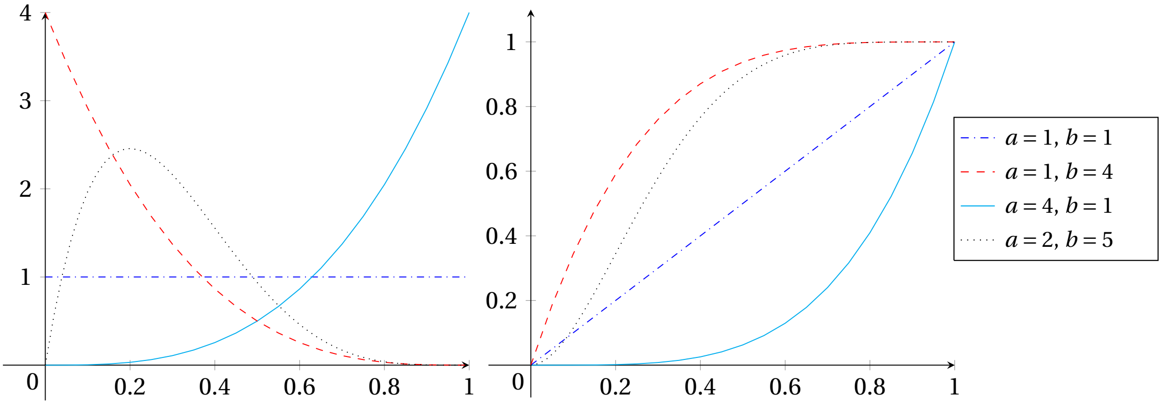
Recall that if $\kappa_i = 0$, the individual is *homo oeconomicus* (fully selfish), and if $\kappa_i = 1$, the individual is called *homo kantiensis* (fully moral) from the name of the German philosopher Immanuel Kant. Indeed, *homo kantiensis* individuals fully endorse the first formulation of Kant (1870) categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law.".

As discussed in Chapter 2, there is substantial theoretical and empirical evidence suggesting that individuals are moral. Alger and Weibull (2013, 2016) recently demonstrated that a particular kind of *homo moralis* arises endogenously as the most favored preference by evolution in a model of preference evolution under incomplete information and assortative matching. Previously, Bergstrom (1995) also proved the evolutionary stability of a "semi-Kantian" utility function (a *homo moralis* with morality coefficient one half) in the special case of symmetric interactions between siblings. Miettinen et al. (2017) have shown that *homo moralis* has a higher explanatory power than other social preferences in a sequential prisoners' dilemma. Finally, we have established that *homo oeconomicus* and *homo kantiensis* individuals can coexist and be favored by evolution in a prisoner's dilemma. We have also shown that heterogeneous populations are more robust than homogeneous populations.

Broadening this diversity, we consider a population of *homo moralis* with different degrees of morality $\kappa_i \in [0,1]$. More precisely, we assume that the individuals' degrees of morality are independently drawn from a given distribution $F(\cdot)$ with density $f(\cdot)$ and support $[0,1]$. In other words, Nature assigns a degree of morality $\kappa_i \in [0,1]$ to each individual by picking from a given distribution. Using the properties of the cumulative distribution function (CDF) $F(\cdot)$ and the probability density function (PDF) $f(\cdot)$, we have:

$$\forall i \in I, \forall k \in [0,1]: \quad P(\kappa_i \leq k) = F(k) = \int_0^k f(t)dt$$

We will often assume that the individuals' degrees of morality follow a Beta distribution, $\kappa \sim \text{Beta}(a,b)$, with the shape parameters $a > 0$ and $b > 0$. The Beta distribution has two major advantages. First, its support is $[0,1]$, hence corresponding to the domain on which the morality coefficient κ is defined. Second, it allows for a large amount of flexibility through its two parameters. Most of the common distributions with a $[0,1]$ support can in fact be represented as a Beta distribution by modulating the two parameters. For instance, when $a = 1$ and $b = 1$, the degrees of morality are uniformly distributed, i.e. $\text{Beta}(1,1) \sim \text{U}(0,1)$. Some examples of Beta distributions are illustrated in Figure 3.3.



(a) Probability density function **(b)** Cumulative distribution function

Figure 3.3 – Distribution of degrees of morality in the population depending on Beta distribution shape parameters

The mean of a Beta distribution is $a/(a+b)$. Thus, an increase in a increases the expected degree of morality while an increase in b decreases the expected degree of morality in the population. Finally, the CDF of the Beta distribution, noted $F_{\beta(a,b)}(\cdot)$ follows:

$$F_{\beta(a,b)}(x) = \frac{\int_0^x t^{a-1}(1-t)^{1-b}dt}{\int_0^1 t^{a-1}(1-t)^{1-b}dt}$$

In particular, we have:

$$\forall x \in [0,1], \quad F_{\beta(1,b)}(x) = 1-(1-x)^b$$
$$\forall x \in [0,1], \quad F_{\beta(a,1)}(x) = x^a$$

Having defined all the ingredients of our model, we carry on with the analysis.

3.3 Why do some individuals cooperate in a social dilemma?

We first analyze why some individuals cooperate and others do not in a social dilemma (Section 3.3.1). Then, we derive the implications for the average level of cooperation in a population (Section 3.3.2). Finally, we discuss how peer pressure influences the level of cooperation (Section 3.3.3).

3.3.1 Individual cooperation

For a given cooperation share $\bar{x}$, when cooperating, *homo moralis* individuals get utility:

$$u_{\kappa_i}(C,\bar{x}) = (1-\kappa_i)\cdot\pi_i(C,\bar{x}) + \kappa_i\cdot\pi_i(C,1)$$

On the other hand, when they defect, they get:

$$u_{\kappa_i}(D,\bar{x}) = (1-\kappa_i)\cdot\pi_i(D,\bar{x}) + \kappa_i\cdot\pi_i(D,0)$$

Individuals cooperate when their utility from cooperating is higher than from defecting, i.e. when $u_{\kappa_i}(C,\bar{x}) - u_{\kappa_i}(D,\bar{x}) \geq 0$. We have:

$$u_{\kappa_i}(C,\bar{x}) - u_{\kappa_i}(D,\bar{x}) = (1-\kappa_i)\cdot[\pi_i(C,\bar{x}) - \pi_i(D,\bar{x})] + \kappa_i\cdot[\pi_i(C,1) - \pi_i(D,0)]$$

Rewriting this equation with the individual cost of cooperating $IC_i(\bar{x})$ (Definition 10) and the social benefit SB_i (Definition 11):

$$u_{\kappa_i}(C,\bar{x}) - u_{\kappa_i}(D,\bar{x}) = -(1-\kappa_i)\cdot IC_i(\bar{x}) + \kappa_i\cdot SB_i \tag{3.1}$$

For a selfish *homo oeconomicus* ($\kappa_i = 0$), we have $u_0(C,\bar{x}) - u_0(D,\bar{x}) = -IC_i(\bar{x}) < 0$. Thus, and this comes as no surprise, *homo oeconomicus* always defects since it is costly to cooperate. On the other hand, for a fully-moral *homo kantiensis* ($\kappa_i = 1$), we have $u_1(C,\bar{x}) - u_1(D,\bar{x}) = SB_i > 0$. Hence, *homo kantiensis* always cooperates because individuals of this type only value what is best for society. More generally, *homo moralis* cooperates when her degree of morality is high enough:

Theorem 4. *For a given cooperation share $\bar{x} \in [0,1]$, a homo moralis cooperates if and only if her degree of morality κ_i is greater than the threshold $\kappa_i^0(\bar{x})$ with:*

$$\kappa_i^0(\bar{x}) = \frac{IC_i(\bar{x})}{IC_i(\bar{x}) + SB_i} \quad \in (0,1)$$

Proof. In Appendix C.1 □

Note that since payoffs can be different for each individual, so can the individual threshold degree of morality $\kappa_i^0(\bar{x})$. Thus, two individuals with the same degree of morality could act differently depending on their incentive to defect and on how they benefit from a situation of full cooperation. Moreover, the threshold can evolve with the cooperation share $\bar{x}$. Hence, one individual could cooperate or defect depending on her peers' behavior. We will discuss this feature in more detail in Section 3.3.3.

The degree of morality necessary for cooperation can be low if the social benefit is high in comparison with the individual cost of cooperating: if $SB_i >> IC_i(\bar{x})$ then $\kappa_i^0(\bar{x}) \approx 0$. In other words, if the individuals think that the externality associated with defection is high, they will be more inclined to cooperate. On the other hand, when the social benefit is low in comparison with the cooperation cost, only individuals with high degrees of morality will cooperate: if $SB_i << IC_i(\bar{x})$ then $\kappa_i^0(\bar{x}) \approx 1$. This observation is in line with our expectations. For instance, individuals are more likely to avoid littering and to recycle when this action requires limited effort (garbage at proximity, process easy to understand, etc.), i.e. when their individual cost of cooperating is low. Similarly, individuals are more likely to support local initiatives to preserve Nature because their social benefit is higher.

3.3.2 Level of cooperation in the population

We have established that individuals cooperate when their degree of morality is high enough. We now analyze what the equilibrium cooperation share $\bar{x}^*$ in the population is.

From Theorem 4, the individual threshold for cooperation is $\kappa_i^0(\bar{x}) = IC_i(\bar{x})/(IC_i(\bar{x})+SB_i)$ for a given cooperation share $\bar{x}$. Since the degrees of morality are independently drawn from a given distribution $\kappa \sim$ F(.), the probability of cooperating for each individual $i \in I$ is the probability that her individual degree of morality is greater than her threshold ($\kappa_i \geq \kappa_i^0(\bar{x})$). We have for all $i \in I$: $\Pr[x_i = 1] = 1 - F(\kappa_i^0(\bar{x}))$. Consequently, the (expected equilibrium cooperation share satisfies:

$$\bar{x}^* = \int_{i\in I} \Pr[x_i = 1]\, d\mu = 1 - \int_{i\in I} F(\kappa_i^0(\bar{x}^*))\, d\mu$$

We have the following Proposition:

Proposition 5. *Consider a population of homo moralis involved in a social dilemma such that the degrees of morality are independently drawn from the distribution $F(\cdot)$. There exists an equilibrium cooperation share $\bar{x}^* \in [0,1]$ such that $\bar{x}^* = 1 - \int_{i\in I} F(\kappa_i^0(\bar{x}^*))\, d\mu$.*

Proof. In Appendix C.2. □

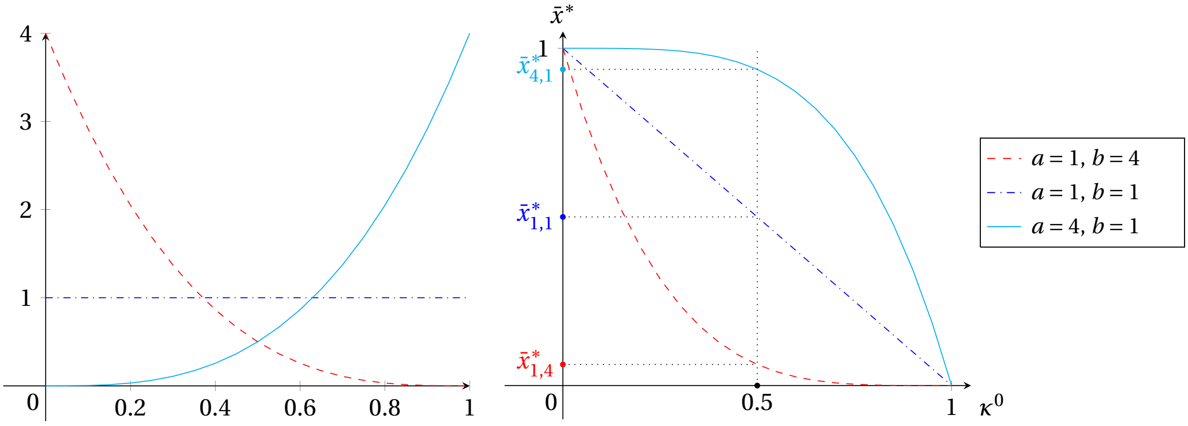
The Proposition proves the existence of an equilibrium cooperation share. Although this might seem obvious, the Proposition is only true because we assumed that the individuals' payoffs are continuous in $\bar{x}$. If this was not the case, there could be a situation where a little group of individuals would always want to deviate so that there does not exist an equilibrium. Moreover, although the Proposition informs about the existence of an equilibrium cooperation share, it tells nothing about its uniqueness. Indeed, there could be more than one equilibrium as we will observe in Section 3.3.3.

For illustrative purposes, we will study in the following the case of a uniform cost-benefit structure (Definition 12): the individual cost and social benefit are the same for all individuals in the population and independent of the cooperation share $\bar{x}$, i.e. for all $i \in I$ and $\bar{x} \in [0,1]$, $IC_i(\bar{x}) = IC$ and $SB_i = SB$. In this context, we have for all $i \in I$ $\kappa_i^0 = \kappa^0 = IC/(IC+SB)$ and all individuals with $\kappa_i \geq \kappa^0$ cooperate. Consequently, the level of cooperation in the population is:

$$\bar{x}^* = 1 - F(\kappa^0)$$

Suppose that the population is homogeneous, i.e. all individuals have the same degree of morality κ. Then, they all make the same decision, i.e. they all defect if $\kappa < \kappa^0$ or they all cooperate if $\kappa > \kappa^0$. Hence, allowing for diversity of preferences is necessary to be in line with the empirically observed diversity of behaviors. For instance, in a population with a share λ of *homo kantiensis* and $(1-\lambda)$ of *homo oeconomicus*, the cooperation share $\bar{x}^*$ is equal to λ since all *homo kantiensis* cooperate and all *homo oeconomicus* defect.

More generally, if the individuals' degrees of morality are independently drawn from a Beta distribution, then the level of cooperation in the population is $\bar{x}^* = 1 - F_{\beta(a,b)}(\kappa^0)$. For example, if the degrees of morality are uniformly distributed (i.e. $a = 1, b = 1$), the cooperation share satisfies: $\bar{x}^* = 1 - \kappa^0 = \frac{SB}{IC+SB}$. If the distribution favors low degrees of morality (e.g. $a = 1, b = 4$), the cooperation share decreases $\bar{x}^* = [SB/(IC+SB)]^4$. Reciprocally, if the distribution favors high degrees of morality (e.g. $a = 4, b = 1$), the cooperation share increases $\bar{x}^* = 1 - [IC/(IC+SB)]^4$. These examples are illustrated in Figure 3.4.



(a) Probability density function

(b) Level of cooperation

Figure 3.4 – Level of cooperation in the population in function of the threshold degree of morality for various distributions of the degrees of morality

Going back to our example on recycling, our model can explain the various recycling rates observed in different countries via two channels. First, one country could have more moral individuals than another. As discussed in Chapter 2, the prevalence of moral individuals is highly dependent on the assortment structure in the population, i.e. on homophily, which is affected by cultural and geographic conditions among others (McPherson et al., 2001; Zhou, 2011). Second, the threshold degree of morality for cooperation κ^0 can be lower in one country than in another, due to a lower individual cost or a higher social benefit. For instance, designing a clear and understandable recycling process with adequate infrastructure or implementing punishment can decrease the incentive to defect. We will discuss the role of policy makers in more detail in Section 3.6. However, the story depicted here is still incomplete. In the following, we will examine how the level of cooperation is influenced by the social structure.

3.3.3 Peer pressure and social norms

Individuals are more inclined to cooperate if they are in a cooperative environment (Fischbacher et al., 2001; Frey and Meier, 2004; Kocher et al., 2008). As discussed in Chapter 2, Section 2.5.3, imitative behaviors have been observed in a variety of context such as environmental and energy conservation (Goldstein et al., 2008; Allcott, 2011) and sustainable food consumption (Vermeir and Verbeke, 2006). Acting as others allows individuals to be accepted in a group and thus gain from network interactions.

Suppose that individuals gain g when sharing the same strategy with others. When they cooperate, the gain is $g\bar{x}$ while they gain $g(1-\bar{x})$ when they defect. We can rewrite the material payoff $\pi(x_i,\bar{x})$ accounting for network interactions:

$$\pi(x_i,\bar{x}) \quad \text{becomes} \quad \pi(x_i,\bar{x}) + g[\bar{x}x_i + (1-\bar{x})(1-x_i)]$$

Then, the additional individual cost ($IC_i(\bar{x}) = \pi_i(0,\bar{x}) - \pi_i(1,\bar{x})$) due to network interactions is $g(1-2\bar{x})$ while the social benefit ($SB_i = \pi_i(1,1) - \pi_i(0,0)$) is unchanged. Suppose that the individual cost and the social benefit are independent of the individuals, i.e. for all $i \in I$ $IC_i(\bar{x}) = IC(\bar{x})$ and $SB_i = SB$. Then, assuming that the individual cost depends on the cooperation share only because of network interactions, we can write: $IC_i(\bar{x}) = IC + g(1-2\bar{x})$.

In this setting, the threshold degree of morality for cooperation is (Theorem 4):

$$\kappa_i^0(\bar{x}) = \frac{IC + g(1-2\bar{x})}{IC + g(1-2\bar{x}) + SB}$$

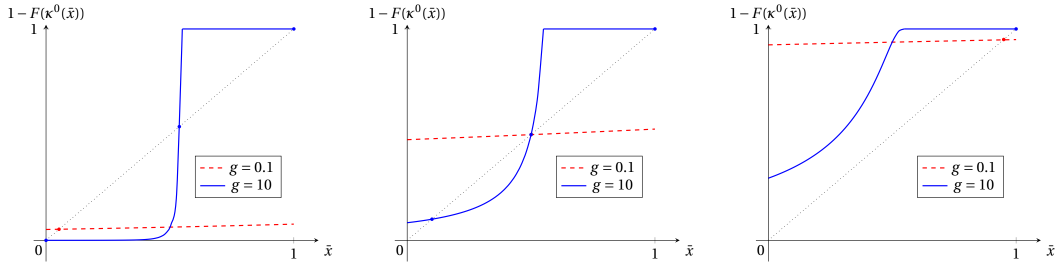
And the level of cooperation in the population satisfies $\bar{x}^* = 1 - F(\kappa^0(\bar{x}^*))$.

We illustrate how peer pressure influences the level of cooperation looking at a few examples. Let $IC = 1$ and $SB = 1$, the threshold degree of morality for cooperation is then $\kappa_i^0(\bar{x}) = (1 + g(1-2\bar{x}))/(2 + g(1-2\bar{x}))$. We analyze two cases. First, we suppose that the gain from network interactions is relatively low $g = 0.1$. Second, we suppose that the gain from network interactions is relatively high $g = 10$. Note that in this case, when the cooperation share is greater than $\bar{x} = 0.55$, the individual cost of cooperating becomes negative $IC(\bar{x}) \leq 0$. Thus, we are no longer in a social dilemma and all individuals are better-off by cooperating. In other words, the threshold degree of morality is zero. We look at three different distributions of the degrees of morality (see Figure 3.3 for a representation of their cumulative and probability density functions):

(a) The distribution favors low degrees of morality: $a = 1, b = 4$. The level of cooperation then satisfies $\bar{x}^* = [1/(2 + g(1 - 2\bar{x}^*))]^4$. When the gain from network interactions is relatively small $g = 0.1$, there is one equilibrium cooperation share $\bar{x}^* = 0.0525$. The level of cooperation is low because of the shape of the distribution. On the other hand, when the gain from network interactions is relatively high $g = 10$, there are three possible equilibrium cooperation shares: the first one is close to zero cooperation, the second is achieved for $\bar{x}^* \approx 0.54$ and in the last one there is full cooperation. Since the peer pressure is much stronger than the incentive to defect (without network effects) and than the social benefit, it drives all individuals to defect, all individuals to cooperate or to the intermediate situation (see Figure 3.5a).

(b) The degrees of morality are uniformly distributed: $a = 1, b = 1$. The level of cooperation then satisfies $\bar{x}^* = 1/(2 + g(1 - 2\bar{x}^*))$. When $g = 0.1$, the unique equilibrium cooperation share is $\bar{x}^* = 0.5$. When $g = 10$, we still have thee potential equilibria: low cooperation $\bar{x}^* = 0.1$, medium cooperation $\bar{x}^* = 0.5$ and full cooperation (see Figure 3.5b).

(c) The distribution favors high degrees of morality: $a = 4, b = 1$. The level of cooperation then satisfies $\bar{x}^* = 1 - [(1 + g(1 - 2\bar{x}^*))/(2 + g(1 - 2\bar{x}^*))]^4$. Due to the shape of the distribution, there is a unique equilibrium cooperation share in both cases: $\bar{x}^* \approx 0.95$ when $g = 0.1$ and full cooperation when $g = 10$ (see Figure 3.5c).



(a) Lowly-moral population ($a = 1, b = 4$)

(b) Midly-moral population ($a = 1, b = 1$)

(c) Highly-moral population ($a = 4, b = 1$)

Figure 3.5 – Influence of peer pressure on the level of cooperation in the population for various distributions of degrees of morality

These examples first show that peer pressure has a strong influence on the level of cooperation. Second, peer pressure can also lead to the emergence of multiple equilibria. Note that in Figures 3.5a and 3.5b for strong network gain ($g = 10$), the intermediate cooperation equilibrium is unstable. When there is a little deviation to the right, more individuals are willing to cooperate than the actual cooperation share $(1 - F(\kappa^0(\bar{x})) > \bar{x})$. Conversely, when there is a little deviation to the left, less individuals are willing to cooperate than the actual cooperation share $(1 - F(\kappa^0(\bar{x})) < \bar{x})$. Thus, strong peer pressure favors either a low or a high cooperation share. Even if the distribution of morality, individual cost (without network gain) and social benefit are the same, we could observe completely opposite levels of cooperation. This suggests that the level of cooperation is also determined by social and cultural norms and it could reflect a path dependency.

Moreover, we assumed that network gains were proportional to the cooperation share, as if individuals had interactions with the whole population. In reality, individuals interact in smaller groups. This could lead to the emergence of groups of cooperators and groups of defectors. Since the formation of networks is influenced by demography, geography, climate and infrastructure among others, the level of cooperation also depends on the environmental and socio-economic characteristics of countries.

In our framework, we included network interactions as an externality affecting the material payoffs of individuals. Another way to model the higher propensity of individuals to cooperate in a cooperative environment would be to add in individuals' utility a reciprocity component as in Levine (1998) or a preference for conformity as in Akerlof (1997). In this setup, it would be as if individuals had internalized the externality when making decisions.

3.4 The effect of misperception

In the model described so far, individuals have perfect knowledge of the average cooperation share in the population $\bar{x}$, their individual cost $IC_i(\bar{x})$, and social benefit SB_i when choosing their strategy. However, this feature is not realistic in all cases and introducing individuals' perception can have major consequences for the model. In this section, we discuss three cases of misperception in our model and the ensuing consequences.

3.4.1 Perceived individual cost

A first case of misperception occurs when the perceived individual cost (noted $\vartheta_i(IC_i(\bar{x}))$) is different from the actual individual cost $(IC_i = \pi_i(D,\bar{x}) - \pi_i(C,\bar{x}))$ expressed in payoff terms. In this setting, the threshold degree of morality for cooperation becomes:

$$\kappa_i^0(\bar{x}) = \frac{\vartheta_i(IC_i(\bar{x}))}{\vartheta_i(IC_i(\bar{x})) + SB_i}$$

Consequently, when the perceived individual cost is greater than the actual individual cost, i.e. when $\vartheta_i(IC_i(\bar{x})) > IC_i(\bar{x})$, the threshold degree of morality for cooperation increases and individuals have less incentives to cooperate. For instance, individuals tend to be attached to private cars and to have a negative image of bus, which hinders the use of public transport (Beirão and Cabral, 2007).

Another example of perceived individual cost is linked to switching costs among strategies. If individuals have to make a decision without being assigned a strategy *a priori*, then they have the same decision cost no matter whether they cooperate or defect, i.e. we are in the case of section 3.3.1. However, if individuals are assigned to a strategy by default or if they are used to act in a given manner, then it requires effort to switch. Individuals might misjudge the effort needed. For instance, if defection is the default strategy (or action done in the past), then the perceived individual cost can be represented by $\vartheta_i(IC_i(\bar{x})) = IC_i(\bar{x}) + e_i$, with e_i being the (perceived) individual effort for switching strategy. Thus, the threshold degree of morality for cooperation increases and individuals have less incentives to cooperate:

$$\frac{\partial \kappa_i^0(\bar{x})}{\partial e_i} = \frac{SB_i}{(IC_i(\bar{x}) + e_i + SB_i)^2} > 0$$

On the other hand, if individuals are assigned to the cooperation strategy by default, then the perceived individual cost is $\vartheta_i(IC_i(\bar{x})) = IC_i(\bar{x}) - e_i$. Hence, the threshold degree of morality for cooperation decreases and individuals have more incentives to cooperate:

$$\frac{\partial \kappa_i^0(\bar{x})}{\partial e_i} = -\frac{SB_i}{(IC_i(\bar{x}) - e_i + SB_i)^2} < 0$$

Note that even a selfish *homo oeconomicus* could cooperate if $IC_i(\bar{x}) < e_i$. Indeed, in this case the perceived individual cost is negative $\vartheta_i(IC_i(\bar{x})) < 0$ and the individual has more incentives to cooperate.

In both cases the average level of cooperation in the population will be altered based on the default option among the population. Specifically, in the case of uniform cost-benefit (Definition 12), we have: $\bar{x}^* = 1 - F(\kappa^0)$. Since the CDF $F(\cdot)$ is increasing, the level of cooperation in the population will be higher when the default option is cooperation and lower when the default option is defection. We will illustrate this observation in Section 3.5.1. Therefore, the problem framing and path dependency matters and could have strong implications on the level of cooperation in the population. This underlines the central role of education as a policy instrument for influencing individuals' behaviors (see Section 3.6).

3.4.2 Perceived social benefit

Another case of misperception occurs at the level of the social benefit ($SB_i = \pi_i(C,1) - \pi_i(D,0)$). For a given cooperation share $\bar{x}$, at the moment of the individuals' decision making, the situations "everybody cooperates" and "everybody defects" are hypothetical, i.e. individuals have to guess what the social benefit is. In some cases, this could be quite challenging. For example, imagine that an individual decides to purchase a new car, having the choice between a conventional fuel-engine or an electric car. Electric mobility coupled with a low-carbon electricity mix emits less greenhouse gas emissions and could decrease urban air pollution, but the production of electric vehicles raises concerns on water pollution and materials depletion (Faria et al., 2012, 2013; Hawkins et al., 2013). Furthermore, there are still many uncertainties associated with climate change impacts, and little is known on the climate impacts on catastrophic events, health or biodiversity (Pindyck, 2013). Hence, each individual has to evaluate the social benefit in the best of her knowledge, pondering different objectives. This evaluation is sensitive to her education and her awareness of the problem, and will thus differ from the "true" social benefit.

For each individual, let $v_i(SB_i)$ be the perceived social benefit of individual i. Then, *homo moralis* cooperates if and only if (Equation 3.1):

$$\kappa_i \cdot v_i(SB_i) \geq (1-\kappa_i) \cdot IC_i(\bar{x})$$

Now suppose that the perceived social benefit is negative, i.e. $v_i(SB_i) < 0$. This means that the individual thinks that her payoff when everybody cooperates is lower than her payoff when everybody defects. Then, we are no longer in a social dilemma, and no matter her degree of morality, the individual defects. Thus, even the fully-moral *homo kantiensis* defects.

More generally, if an individual underestimates the social benefit, i.e. if $v_i(SB_i) < SB_i$, then the threshold degree of morality for cooperation $\kappa_i^0(\bar{x}) = IC_i(\bar{x})/(IC_i(\bar{x}) + v_i(SB_i))$ increases. In turn, the probability to cooperate $1 - F(\kappa_i^0(\bar{x}))$ decreases. If the whole population underestimates the social benefits, i.e. if for all $i \in [0,1]$ $v_i(SB_i) \leq SB_i$, then the expected cooperation share $\bar{x}^* = 1 - \int_0^1 F(\kappa_i^0(\bar{x}^*))di$ also decreases. We will illustrate this observation in Section 3.5.1.

Consequently, the social awareness of individuals has a significant impact on their actions and in turn on the level of cooperation in the population. Being knowledgeable of the impacts of human actions on Nature can completely switch individuals' behavior. Conversely, underestimating the impacts of environmentally harmful behavior can lead to socially inefficient situations. Raising the population's awareness appears like a necessary action by policy makers and we will further discuss it in Section 3.6.

3.4.3 Perceived level of cooperation

Even with a perfect knowledge of their individual cost and social benefit, individuals may not know how many cooperators there are in the population, i.e. they misperceive the actual level of cooperation in the population ($\bar{x}^*$). This could happen when the externality is not observable (e.g. greenhouse gas emissions) and when the individuals' behavior is their private information (e.g. purchasing or not green electricity). Thus, individuals have to form beliefs $\upsilon_i(\bar{x})$ about the cooperation share $\bar{x}$ gathering information from the behaviors of individuals in their network or from the media.

In this context, *homo moralis* cooperates if $\kappa_i \geq \kappa_i^0(\upsilon_i(\bar{x}))$ with:

$$\kappa_i^0(\upsilon_i(\bar{x})) = \frac{IC_i(\upsilon_i(\bar{x}))}{IC_i(\upsilon_i(\bar{x})) + SB_i}$$

Now suppose that the individual underestimates the cooperation share (by putting too much emphasis on defecting behaviors for instance), i.e. $v_i(\bar{x}) \leq \bar{x}$. In this situation, the individual probability to cooperate can be lower than under perfect perfection if the individual cost of cooperating decreases with $\bar{x}$, for instance under peer pressure or when individuals have a tendency for prosocial conformity (Section 3.3.3). By contrast, individuals exhibiting anti-conformity are more likely to cooperate when they underestimate the level of cooperation in the population.

At the population scale, if individuals underestimate the level of cooperation (e.g. because of negative news) and if they tend to conform to social norms, then the level of cooperation is lower than under perfect information. For instance, Frey and Torgler (2007) have shown that tax compliance decreases with perceived tax evasion. This underlines the importance of the framing of news transmission in society. Insisting only on environmentally-harmful behaviors such as the fraud on emission levels by some firms can deter some individuals from performing environmentally-friendly actions because of their negative perception. On the other hand, putting forward positive actions and cooperative behavior could have a surprisingly positive impact on the environment.

3.5 Applications in environmental and resource economics

In this section, we first illustrate our model and the effects of misperception with an application on green-electricity purchase (Section 3.5.1). Then, we look at the adoption of environmentally-friendly products in a dynamic framework (Section 3.5.2). Finally, we extend the model introducing mixed strategies in the context of resource-use (Section 3.5.3).

3.5.1 Purchasing green electricity

In many countries, electricity providers offer the option to pay a premium to get "green" electricity, i.e. electricity produced with renewable energy sources (e.g. wind, solar or hydraulic). Indeed, there is a large literature showing the high willingness to pay for green electricity (e.g. Roe et al., 2001; Hansla et al., 2008; Sundt and Rehdanz, 2015). A recent study conducted among EPFL students and staff members offers some interesting insights on this unselfish behavior (Detsouli, 2018).[4] First, 85% of respondents are willing to pay more to get green electricity, but only 28% declared they do. This inconsistency suggests that individuals are not aware that this option is available. Actually, most Swiss electricity providers recently changed their standard product so that consumers pay more and get green electricity by default.[5] If individuals wish to change, they need to contact their electricity provider. This observation

[4] This study was part of a semester project supervised by my coauthor Charles Ayoubi and Prof. Foray. I provided occasional feedback and participated in the survey design.

[5] In particular, the "Service Industriel de Lausanne" (SIL) and "Romande Energie" supply most of the Lausanne area and their default product comprises 100% renawables (see http://paysage-electricite.mynewenergy.ch/ for a map of the default electricity mix by Swiss municipalities).

highlights the misperception among individuals. Moreover, about 30% of respondents stated that the share of renewables in their electricity mix is below 15%. But renewables account for about two thirds of the Swiss electricity production (SFOE, 2018). Thus, there is a lack of knowledge even though the vast majority of the respondents is sensitive to the environment and willing to make effort. In the following, we present a simple model to shed some lights on these observations and to illustrate the effects of misperception.

Individuals get the utility $V_i(e_i)$ and pay the cost $\gamma_i(e_i,\bar{x})$ when they consume the quantity e_i of electricity. In addition, they can subscribe to an option to get green electricity (i.e. cooperate) by paying an additional amount $\gamma_g(e_i,\bar{x})$, where we assume that $\gamma_g : \mathbb{R}\times[0,1]\to\mathbb{R}$ decreases when the cooperation share $\bar{x}$ increases to represent the decrease in price of renewables with their adoption. Note that the potential electricity price change due to renewables integration is already included in the cost γ_i. We assume that individuals' demand for electricity e_i is the same whether they subscribe to the options or not, either because they need e_i in their everyday activities or because γ_g is small in comparison with γ_i.[6] The production of conventional fossil-fuel electricity emits greenhouse gases contributing to climate change while a greater share of renewables in the electricity mix could increase land and material use (Gagnon et al., 2002). The associated net externality is called $\xi(\bar{x})$. Since the population is large, each individual choice has no effect on the externality. For each cooperation share $\bar{x}$, the payoff to purchase green electricity (C) or not (D) is:

$$\pi_i(C,\bar{x}) = V_i(e_i) - \gamma_i(e_i,\bar{x}) - \gamma_g(e_i,\bar{x}) - \xi(\bar{x})$$
$$\pi_i(D,\bar{x}) = V_i(e_i) - \gamma_i(e_i,\bar{x}) - \xi(\bar{x})$$

Hence, the individual cost of purchasing green electricity is $IC_i(\bar{x}) = \gamma_g(e_i,\bar{x})$ and the social benefit if everybody cooperates is $SB_i = \gamma_i(e_i,0) - \gamma_i(e_i,1) - \gamma_g(e_i,1) + \xi(0) - \xi(1)$. For simplicity, we will assume that the additional cost of purchasing green electricity γ_g does not depend on the demand for electricity and that the change in electricity cost $(\gamma_i(e_i,0) - \gamma_i(e_i,1))$ is negligible compared to the gain of reducing the externality $(\xi(0) - \xi(1))$. Thus, the individual cost of purchasing green electricity and the social benefit are independent of the individual, i.e. $IC_i(\bar{x}) = IC(\bar{x})$ and $SB_i = SB$. These simplifying assumptions could be relaxed without qualitatively affecting our findings.

Under perfect perception, *homo moralis* decides to purchase green electricity if:

$$\kappa_i SB \geq (1-\kappa_i) IC(\bar{x}) \qquad \Leftrightarrow \qquad \kappa_i \geq \kappa^0(\bar{x}) = \frac{IC(\bar{x})}{IC(\bar{x}) + SB}$$

The level of cooperation satisfies $\bar{x}^* = 1 - F(\kappa^0(\bar{x}^*))$. Let F be a Beta distribution favoring low degree of morality: $a = 1, b = 4$. Then, the level of cooperation satisfies:

$$\bar{x}^* = \left(\frac{SB}{IC(\bar{x}) + SB}\right)^4$$

[6] In Detsouli (2018)'s study, 90% of respondents indicated that they do not know their electricity consumption.

For instance, when $\gamma_g(\bar{x}) = 0.15 - 0.1\bar{x}$ and $SB = 1$, we obtain $\bar{x}^* \approx 0.75$ (see Figure 3.6). The level of cooperation is high even though the population is lowly-moral because the individual cost of purchasing green electricity is small in comparison with the benefits of mitigating climate change.

We study now how misperception can affect the level of cooperation:

1. Perceived individual cost (PIC): we assume that the default option is to purchase the standard (dirty) electricity mix. In order to get the green electricity mix, individuals should send a letter. This effort translates into a switching cost e, assumed the same for all individuals for simplicity. Thus, the payoff to purchase green electricity becomes: $\pi_i(C,\bar{x}) = V_i(e_i) - \gamma_i(e_i,\bar{x}) - \gamma_g(\bar{x}) - e - \xi(\bar{x})$. In turn, the individual cost is $IC(\bar{x}) = \gamma_g(\bar{x}) + e$ and the social benefit $SB = 1 - e$ and individuals decide to purchase green electricity if: $\kappa_i \geq \kappa^0_{PIC}(\bar{x}) = (\gamma_g(\bar{x}) + e)/(\gamma_g(\bar{x}) + 1) > \kappa^0(\bar{x})$. Because the threshold degree of morality increases, the cooperation share decreases. For example with $e = 0.1$, we get $\bar{x}^*_{PIC} \approx 0.44$ (see Figure 3.6).
2. Perceived social benefit (PSB): we assume that there is a lack of awareness regarding climate change, such that the perceived social benefit is $v(SB) = 0.3$, the same for all individuals for simplicity. For instance, individuals could be more worried about the implementation of wind turbines close to their home than about climate change. Individuals decide to purchase green electricity if $\kappa_i \geq \kappa^0_{PSB} = (\gamma_g(\bar{x}))/(\gamma_g(\bar{x}) + 0.3) > \kappa^0(\bar{x})$. Because the threshold degree of morality increases, the cooperation share sharply decreases: $\bar{x}^*_{PSB} \approx 0.25$ (see Figure 3.6).
3. Perceived level of cooperation (P$\bar{x}$): we assume that individuals underestimate the cooperation share such that $v(\bar{x}) = \bar{x}^3$ for all individuals. This affects the individual cost $IC(\bar{x}) = \gamma_g(\bar{x}^3)$. In turn, the threshold degree of morality increases $\kappa_i \geq \kappa^0_{P\bar{x}}(\bar{x}) = \gamma_g(\bar{x}^3)/(\gamma_g(\bar{x}^3) + 1) > \kappa^0(\bar{x})$ because γ_g is a decreasing function. Hence, the cooperation share decreases: $\bar{x}^*_{P\bar{x}} \approx 0.62$ (see Figure 3.6).

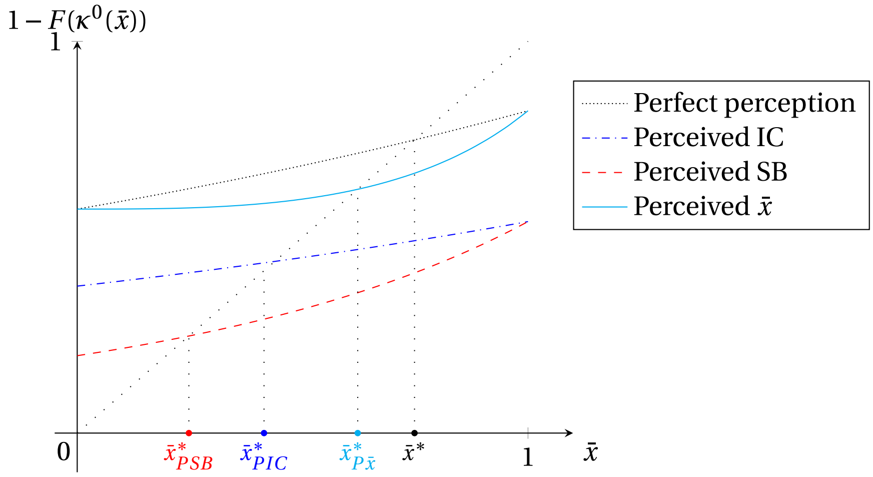


Figure 3.6 – Effects of misperception on the share of individuals purchasing green electricity

Our simple model allows to explain why we can observe high levels of cooperation even in lowly-moral population and even though individuals' behaviors have a negligible impact on the social benefit. Figure 3.6 also illustrates that individuals' belief and misperception can have a drastic impact. In our example when individuals undervalue the externality (perceived social benefit), about 50% of individuals do not purchase green electricity anymore. In this case, financial incentives such as taxes and subsidies would have limited impacts, and policy-makers should rely on education instead. Moreover, the problem framing matters: when defection is the default strategy, less than half of the individuals purchase green electricity. By contrast, if we had assumed that the default strategy was cooperation, we would end-up in a case of full-cooperation. Making decisions can be challenging, specially when one has to weigh several criteria or when the outcome is uncertain. The efforts required to decide and to act thus favor the *status quo*. However, in our context, it seems that individuals are just not aware of their options. In any case, opt in/opt out is an interesting and cheap nudge for policy makers to promote the socially-efficient strategy. We will further discuss the policy implications in Section 3.6.

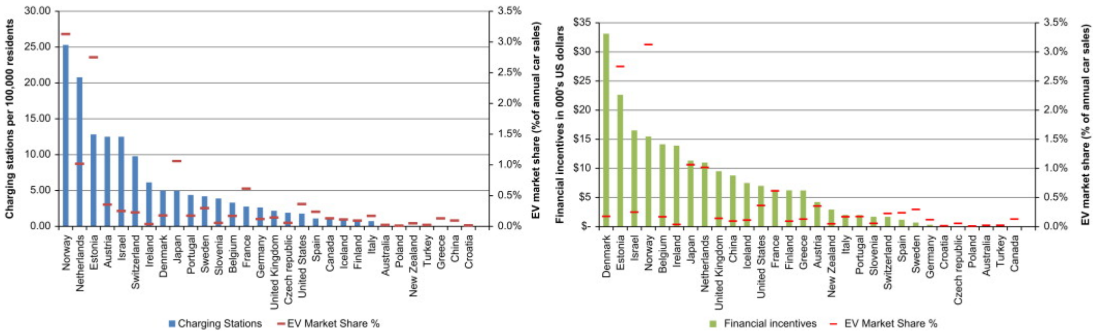
3.5.2 Adoption of environmentally-friendly technologies

In this section, we discuss the adoption of environmentally-friendly technologies taking the example of electric vehicles (EV). Electric mobility could decrease environmental and social externalities by reducing greenhouse gas emissions, urban air pollution and noise (Faria et al., 2012, 2013; Hawkins et al., 2013). Thus, the European Commission has implemented in 2009 the Clean Vehicles Directive to promote environmental-friendly vehicles.[7] However, an ex-post evaluation of the Directive revealed a limited effectiveness and efficiency (Brannigan et al., 2018). There are several barriers to the adoption of EV such as their price, the lack of charging infrastructure and the performance of batteries (Egbue and Long, 2012). In turn, the adoption rates of EV vary between countries and could be partly explained by financial incentives and the charging infrastructure (or lack thereof) (Sierzchula et al., 2014). Nevertheless, financial incentives are not always successful. For instance, Denmark, Israel, Belgium or the United Kingdom had relatively high financial incentives in 2012, but their EV market share was low (see Figure 3.7). In the following, we extend our basic model introducing a dynamic framework to better understand these observations.

We assume that each year, a share α of individuals in the population decides to purchase a new vehicle, having the choice between a conventional fuel-engine (defect) or an electric car (cooperate).[8] Electric vehicles are more expensive, but their cost decreases with their adoption which reflects learning opportunities. This means that the individual cost of cooperating $IC(\bar{x}_t)$ decreases with the EV share in the total fleet $\bar{x}_t$ in year $t \in \mathbb{N}$. We assume that $IC(\bar{x}_t) =$

[7] Directive 2009/33/EC of the European Parliament and of the Council of 23 April 2009 on the promotion of clean and energy-efficient road transport vehicles: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009L0033

[8] The share α represents the share of individuals that decide to purchase a new vehicle and have the financial capacity to buy an electric car.



(a) Number of charging stations

(b) Financial incentives

Figure 3.7 – Share of electric vehicles, financial incentives and number of charging stations in 2012 per country. Source: Reprinted from Sierzchula et al. (2014). The influence of financial incentives and other socio-economic factors on electric vehicle adoption. *Energy Policy*, 68:183–194, 2014, with permission from Elsevier, License Number: 4523091261282, 2019.

$2.5-3\sqrt{\bar{x}_t}$. Note that when $\bar{x}_t \geq 0.7$, the individual cost of cooperating is negative. Thus, for a high adoption share, it becomes more interesting to purchase an electric vehicle than a conventional fuel-engine car. The EV adoption decreases negative environmental and social externality. We assume that the social benefit is $SB = 1$.

Individuals have *homo moralis* preferences. Thus, when they decide to purchase a new car in year t, they prefer EV if their degree of morality is higher than the threshold: $\kappa^0(\bar{x}_t) = \max[IC(\bar{x}_t)/(IC(\bar{x}_t)+SB), 0]$. Since the individual cost of cooperating decrease with the EV share $\bar{x}_t$, the threshold allowing cooperation also decreases with $\bar{x}_t$ and becomes zero when $\bar{x}_t \geq 0.7$ (Figure 3.8). Indeed, EV is then the cheaper option and is attractive even for *homo oeconomicus.*

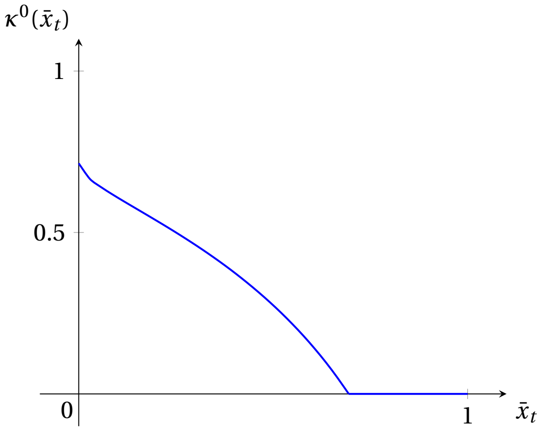


Figure 3.8 – Threshold allowing cooperation in function of EV share in year t

Individuals' degrees of morality are independently drawn from a Beta distribution with $a = 1$ and $b = 4$. Thus, the probability that an individual i purchasing a new car buys an electric vehicle in year t is: $\Pr[x_i = 1] = \min[(SB/(IC(\bar{x}_t)+SB))^4, 1]$. Moreover, the probability that an individual purchases a new car in year t is α. Consequently, the EV share in year $(t+1)$

updates to:

$$\bar{x}_{t+1} = (1-\alpha)\bar{x}_t + \alpha \min[(SB/(IC(\bar{x}_t)+SB))^4, 1]$$

We assume that EV enter the market in year 0, i.e. $\bar{x}_0 = 0$. Then, only individuals with a high degree of morality ($\kappa_i \geq \kappa^0(0) \approx 0.71$) will install the technology. Thus, the EV share at the beginning of year 1 is low, $\bar{x}_1 \approx 0.006$. Thanks to the learning effect, the individual cost of cooperating is lower in year 1 than in year 0. In turn, the threshold allowing cooperation decreases and the EV share will increase. This process goes on until the EV share converges. Let $f : [0,1] \to \mathbb{R}$ and $g : [0,1] \to \mathbb{R}$ the functions such that $\bar{x}_{t+1} = f(\bar{x}_t)$ and $g(x) = f(x) - x$. The zeros of g are the fixed point of $\{\bar{x}_t\}_{t=0}^{\infty}$. The function g admits three zeros: $\bar{x}_\infty^- \approx 0.009$, $\bar{x}^0 \approx 0.628$ and $\bar{x}_\infty^+ = 1$. When $\bar{x}_t < \bar{x}^0$, $\{\bar{x}_t\}_{t=0}^{\infty}$ converges to $\bar{x}_\infty^-$. Conversely, when $\bar{x}_t > \bar{x}^0$, $\{\bar{x}_t\}_{t=0}^{\infty}$ converges to $\bar{x}_\infty^+$ (see Figure 3.9).

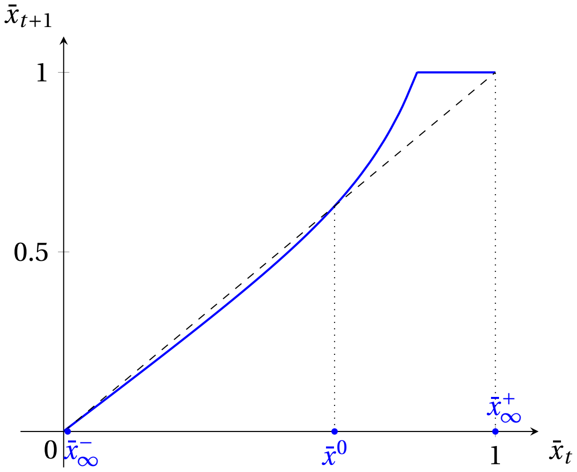


Figure 3.9 – Evolution of EV share due to decreasing adaption cost ($\alpha = 0.1$)

Consequently, the EV share remains very low, reaching the socially-inefficient equilibrium $\bar{x}_\infty^- \approx 0.009$. In our framework, the observed diversity of EV market shares across countries can be explained via several channels. First, the distribution of morality could vary from one country to another. Second, the individual cost of cooperating is affected by regulations (e.g. financial incentives) and by socio-economic and geographic characteristics such as infrastructure development, urbanization and the organization of the territory. Lastly, the social benefit, and in turn the decision to cooperate, depends on individuals' perception.

Furthermore, this illustration shows that financial incentives could be quite ineffective. For instance, if the government offers subsidies decreasing the initial individual cost of cooperating by 20% ($IC(\bar{x}_t) = 2 - 3\sqrt{\bar{x}_t}$), the EV share slightly increases to $\bar{x}_\infty^- \approx 0.024$. Note that in our framework, the individual cost of cooperating does not include only EV financial cost but also others adoption barriers such as the lack of charging infrastructure. Subsidies decreasing $IC(\bar{x}_t)$ by 20% could thus represent a much higher share of the EV price. Therefore, only an appropriate mix of policies encompassing subsidies, investment in infrastructure and education program would allow society to reach full cooperation and escape the technological inertia and lock-in (Cowan and Hultén, 1996).

The adoption of environmentally-friendly technologies actually raise several policy dilemmas.[9] For example, there is a tradeoff between supporting long-term technological alternatives and short-term solutions, which could slow down long term innovation. Policy makers also have conflicting objectives between favoring a competitive environment and promoting one technology to decrease its cost. This could be especially challenging when there exist several technologies that could achieve the same objective. Finally, the emergence of a new technology is associated with uncertainties about its environmental impacts. Knowledge about these impacts could sometimes only be acquired once the technology is already widely adopted.

Finally, we focused in this example on the growth of the EV share due to a decrease in adoption cost, abstracting from technological improvement. It would be possible to also implement a decrease in EV price through time ($IC(\bar{x}, t)$) and an improvement of battery quality lowering the externality ($\xi(\bar{x}, t)$). As a result, the function f would move up at each time period. When the technology is mature-enough, the only remaining equilibrium then is full adoption.

3.5.3 Resource use and sustainable food

Until now, we have assumed that individuals could either defect or cooperate. But in many situations, the choice is not only black or white. For instance, individuals can decide to purchase sustainable food as a share of their total food basket.[10] Empirical evidence reveals that the willingness to pay for sustainable food is positive, heterogeneous (Thilmany et al., 2008; Vecchio, 2013; De Magistris et al., 2015), and highly-dependent on peer pressure (Vermeir and Verbeke, 2006; Adams and Salois, 2010). Since a shift toward sustainable food could improve health, decrease material and energy use, reduce water and soil pollution using less fertilizers and pesticides and preserve biodiversity (Heller and Keoleian, 2003; Roy et al., 2009), many countries aim to promote organic farming and environmental-friendly fishing.[11] For example, the European Common Agricultural Plan includes green payment to support sustainable agriculture. Understanding why individuals are willing to purchase more expensive sustainable food is thus crucial to implement effective policies. In the following, we enrich our basic model allowing for mixed strategies in order to better apprehend individuals' food choices. This extension is inspired by Alger and Weibull (2017)'s work. In a section of their paper, they study the behavior of *homo moralis* in a model of two goods, one environmentally neutral and the other environmentally harmful.

[9] See Foray and Grübler (1996) for a detailed description of technology policy dilemmas for environmental issues.

[10] We understand sustainable food in a broad term, encompassing organic and local products, certified fish, and diet changes favoring vegetables and fruits among others.

[11] See for instance the detailed European Action Plan on organic farming https://ec.europa.eu/agriculture/organic/eu-policy/european-action-plan_en and EU fisheries policies https://ec.europa.eu/fisheries/reform/.

We call $x_i \in [0,1]$ the share of sustainable food in the total food consumption of individual i. When $x_i = 0$, the individual does not consume sustainable food, i.e. she "defects". When $x_i = 1$, the individual only consumes sustainable food, i.e. she "cooperates". The average consumption of sustainable food in the population $\bar{x}$ is thus called the cooperation share. We assume atomicity, i.e. a change in sustainable food consumption of individual i does not affect the aggregate consumption of sustainable food:

Property 7 (Atomicity)**.** The average level of cooperation in the population is unaffected by the action of a single individual:

$$\forall i \in I, \forall \bar{x} \in [0,1], \forall x_i \in [0,1]: \quad \partial \bar{x} / \partial x_i = 0$$

Consuming more sustainable food means that individuals have to make some sacrifices such as consuming less meat or giving up exotic vegetables and fruits. On the other hand, it could procure additional satisfaction thanks to better-quality products or improved health. We represent these effects in the hedonic utility $V_i(x_i)$ where we assume that $V_i : [0,1] \to \mathbb{R}$ is continuous and differentiable for all individuals. Moreover, we suppose that purchasing a share x_i of sustainable food requires an additional spending $\gamma_s x_i$ due to loss from trade and costly environmental-friendly production techniques. Finally, the consumption of sustainable food is associated with better environmental conditions, i.e. it decreases the negative externality $\xi(\bar{x})$, where $\xi : [0,1] \to \mathbb{R}$ is assumed continuous, strictly decreasing and differentiable. Putting it all together, the individuals' payoff $\pi_i : [0,1]^2 \to \mathbb{R}$ is:

$$\pi_i(x_i, \bar{x}) = V_i(x_i) - \gamma_s x_i - \xi(\bar{x})$$

Individuals have *homo moralis* preferences:

$$\begin{aligned} u_{\kappa_i}(x_i, \bar{x}) &= (1-\kappa_i) \cdot \pi(x_i, \bar{x}) + \kappa_i \cdot \pi(x_i, x_i) \\ &= V_i(x_i) - \gamma_s x_i - (1-\kappa_i)\xi(\bar{x}) - \kappa_i \xi(x_i) \end{aligned}$$

Hence, for a given $\bar{x}$, they solve the following maximizing problem:

$$x_i^* \in \arg \max_{x_i \in [0,1]} \{V_i(x_i) - \gamma_s x_i - (1-\kappa_i)\xi(\bar{x}) - \kappa_i \xi(x_i)\}$$

If an individual defects, then for all $x_i \in [0,1]$, $V_i(0) - \kappa_i \xi(0) \geq V_i(x_i) - \gamma_s x_i - \kappa_i \xi(x_i)$. Thus, her degree of morality satisfies:

$$\kappa_i \leq \frac{V_i(0) - V_i(x_i) + \gamma_s x_i}{\xi(0) - \xi(x_i)} \qquad \forall x_i \in (0,1]$$

Conversely, if an individual cooperates, then for all $x_i \in [0,1]$, $V_i(1) - \gamma_s - \kappa_i \xi(1) \geq V_i(x_i) -$

$\gamma_s x_i - \kappa_i \xi(x_i)$ so that that her degree of morality satisfies:

$$\kappa_i \geq \frac{V_i(x_i) - V_i(1) + \gamma_s(1 - x_i)}{\xi(x_i) - \xi(1)} \qquad \forall x_i \in [0,1) \tag{3.2}$$

In particular, if for all $x_i \in [0,1]$ $V_i(x_i) + \gamma_s(1 - x_i) \leq V_i(1)$ then the individual cooperates no matter her degree of morality since the right-hand side of Equation 3.2 is negative. This could be the case if the individual loves vegetarian food and pays close attention to her health.

When a mixed strategy $x_i \in (0,1)$ is optimal, then:

$$\frac{\partial V_i(x_i^*)}{\partial x_i} = \gamma_s + \kappa_i \frac{\partial \xi(x_i^*)}{\partial x_i}$$

For example, with $V(x_i) = 0.5(1 + x_i - x_i^2)$, $\gamma_s = 0.5$ and $\xi_i(\bar{x}) = 1 - \bar{x}$, we obtain $x_i^* = \kappa_i$.

Then, when the individuals degree of morality are independently drawn from a Beta distribution $\kappa \sim \beta(a,b)$, the density of people playing $x_i^* = \kappa_i$ is $f_{\beta(a,b)}(x_i^* = \kappa_i)$. Thus, the expected cooperation share is $\bar{x}^* = \int_{i \in I} x_i^* f_{\beta(a,b)}(x_i^*) dx_i^* = \int_{i \in I} \kappa_i f_{\beta(a,b)}(\kappa_i) d\kappa_i$ and is equal to the mean of the Beta distribution:

$$\bar{x}^* = \frac{a}{a+b}$$

For instance, in a lowly-moral population ($a = 1, b = 4$), we have $\bar{x}^* = 20\%$.

This application illustrates that our basic framework can easily be enriched and that heterogeneous moral populations can explain environmental-friendly behaviors in a variety of contexts. However, our simple model has some substantial limitations. First, we mentioned that peer pressure has a strong influence on food consumption but we did not include it in our example. Introducing network gains is of limited complexity but would not lead to an easy analytical solution. Second, and more importantly, we did not incorporate individuals' wealth and the production side, such that income effects are missing and price effects are incorrectly portrayed. A proper analysis would require a general equilibrium approach. Therefore, this application should be understood as a basis for further research.

Nonetheless, our model offers some interesting insight. In particular, if we focus on local food associated with a decrease in environmental and social externality (e.g. due to less transport or stricter environmental norms), a moral population will have a higher demand for local products than a population of *homo oeconomicus*. This observation is in line with empirical evidence showing that there is too little international trade and too much intranational trade (McCallum, 1995; Wolf, 2000). This "home bias puzzle" is unexplained by trade cost barriers. While most of the trade literature focuses on the production side and border effects (Evans, 2003; Yi, 2010), Caron et al. (2014) argue that the characteristics of individuals' preferences should be taken into account. Consequently, implementing morality in a trade model could help solving part of the puzzle by better representing the consumer side.

3.6 Policy implications

In a model with *homo oeconomicus* agents, individuals are only receptive to financial and regulatory incentives. Introducing heterogeneous moral agents provides new insights on potential policies to promote cooperation. We first discuss the effectiveness of financial instruments in our setting (Section 3.6.1), before looking at the effects of nudges (Section 3.6.2), signaling, learning and education (Section 3.6.3) and urban planning (Section 3.6.4).

3.6.1 On the effectiveness financial instruments

Financial incentives are widely used to promote environmental-friendly behaviors.[12] For instance, subsidies are implemented to enhance energy efficiency in buildings (e.g. insulation, efficient boilers and electric appliances), renewable energy installation (e.g. photovoltaic, solar thermal, heat pump) and clean transport (e.g. electric and hybrid cars). The subsidies can be complemented by carbon taxes on fossil fuels and by feed-in premia and tariffs for green electricity. The European Common Agricultural Plan includes green payments to farmers to support crop diversification, maintenance of permanent grasslands and ecological focus areas. In Switzerland, waste management incorporates garbage collection and rubbish bag taxes, and fines as defined by the Environmental Protection Act. In this section, we analyze the effectiveness of financial incentives in promoting cooperative behaviors in our framework. The instrument is effective when it is pursuing the right goal, i.e. it helps increasing the cooperation share.

Let τ_i be a financial incentive. When τ_i is a (lump-sum) tax or a fine, the individual payoff of defecting $\pi_i(D,\bar{x})$ becomes $(\pi_i(D,\bar{x}) - \tau_i)$. When τ_i is a subsidy, the individual payoff of cooperating $\pi_i(C,\bar{x})$ becomes $(\pi_i(C,\bar{x}) + \tau_i)$. In both cases, the financial incentive decreases the individual cost of cooperating and increases the social benefit: $IC_{\tau i}(\bar{x}) = IC_i(\bar{x}) - \tau_i$ and $SB_{\tau i} = SB_i + \tau_i$, with $IC_i(\bar{x})$ and SB_i the individual cost and social benefit in the absence of incentives. Under perfect perception, *homo moralis* cooperates if and only if:

$$\begin{aligned} & \kappa_i \cdot (SB_i + \tau_i) \geq (1-\kappa_i) \cdot (IC_i(\bar{x}) - \tau_i) \\ \Rightarrow \quad & \tau_i \geq (1-\kappa_i) IC_i(\bar{x}) - \kappa_i SB_i \end{aligned}$$

Note that if individuals are *homo oeconomicus*, the financial incentive is effective only if $\tau_i \geq IC_i(\bar{x})$. By contrast, a moral individual needs a lower incentive to adjust her behavior since she includes the social benefit when making her decision. Yet, the financial incentive is ineffective if it is smaller than the threshold $\tau_i^0 = (1-\kappa_i) IC_i(\bar{x}) - \kappa_i SB_i$. In particular, a Pigovian tax on the externality would not change individuals' behavior in our setting since they all have a negligible impact on the externality (atomicity assumption). As observed in Section 3.5.2, the electric-vehicles market share remains low in some countries despite

[12] The *Climate Policy Database* provides a list of policies related to climate change mitigation worldwide, see http://climatepolicydatabase.org/; the *MURE Database* lists energy efficiency policies in Europe, see http://www.measures-odyssee-mure.eu/.

relatively high financial incentives. Similarly, a carbon tax on gasoline has a limited effect on driving behaviors, resulting in a small improvement of the quality of the environment (Sipes and Mendelsohn, 2001). In our setting, these observations can be explained by the high individual cost of cooperating, due to the lack of charging infrastructures for electric vehicles and public transport options among others.

Moreover, financial incentives should be tailored to each individual to be efficient. Our framework suggests that the optimal financial incentive ($\tau_i^0 = (1-\kappa_i)IC_i(\bar{x}) - \kappa_i SB_i$) depends on the individual cost, social benefit and also degree of morality. However, policy makers cannot easily infer the degrees of morality of each individual. Only behaviors can be observed, and not preferences. Still, it is possible to take advantage of the diversity of morality in the population. For instance, one could imagine a transfer mechanism such that individuals with the highest propensity to cooperate cross-subsidize the ones with the lowest. This could take the form of a voluntary environmental tax, the proceeds being used to fund environmental projects. Furthermore, when the individual cost is decreasing with the cooperation share (for instance when peer pressure is strong, see Section 3.3.3, or in a dynamic settings, see Section 3.5.2), a financial incentive increases the cooperation share via two channels. First, all individuals with $\tau_i \geq \tau_i^0$ directly adjust their behaviors. Second, since the cooperation share increases, the individual cost decreases and so do the threshold degrees of morality. Hence, even a moderate incentive ($\tau_i < IC_i(\bar{x})$) could lead to a situation of full cooperation.

However, these observations only hold true under perfect perception. Indeed, when individuals overestimate the individual cost or when they underestimate the social benefit or the cooperation share, the financial incentives needed to promote cooperation could be excessively high to cover for the misperception. This observation is in line with empirical evidences showing that individuals question the effectiveness of environmental taxes. Public acceptance depends on beliefs about environmental consequences and on concerns about distributional, competitiveness and employment effects (Thalmann, 2004; Kallbekken and Sælen, 2011; Carattini et al., 2017). Thus, policy makers should better communicate on the environmental, social and economic impacts of taxation. They should also be careful on the instrument design and favor progressive designs, e.g. recycling via lump-sum transfers, such that lower income groups do not pay the burden of the tax. Finally, policy makers could improve trust in the society and remedy a low perception of the cooperation share by increasing financial penalties and controls.

Nonetheless, relying on financial incentives could have major drawbacks if their implementation disrupts individuals' perception. For example, an insufficient tax may send the wrong signal that the externality is fully-compensated. In other words, individuals would no longer consider the social benefit when making a decision, as if they were *homo oeconomicus*. Furthermore, empirical evidence suggests that individuals are more selfish when they evolve in a monetary environment. For instance, reminding individuals of money decreases helping behaviors (Vohs et al., 2006) and increases endorsement of social inequality (Caruso et al., 2013). In a famous field-study in day-care centers, Gneezy and Rustichini (2000) have also shown that the introduction of a fine for parents arriving late to collect their children significantly increased the number of late-coming parents. Even worse, the effect was not reversible: after the fine was removed, no reduction occurred. These observations suggest that by putting a price on Nature, individuals could leave the "moral sphere" towards the "economic sphere". As a result, the distribution of morality in the population would shift down and the cooperation share would decrease.

3.6.2 Nudges

Thaler and Sunstein (2008, p. 6) defines nudges as an intervention that *"alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives"*. Nudges have received increasing interest because they are cheap, they do not limit individuals' freedom of choice and they are generally well-accepted by citizens (Hagman et al., 2015; Reisch and Sunstein, 2016).[13] Nudges have also proven effective in a variety of contexts, helping translate the good intentions of individuals into actions (Byerly et al., 2018). For example:

- In a field experiment in a university, the recycling rate of plastic cups has increased from 4% to almost 100% by showing a positive message and by increasing the relative size of the recycling garbage (Cosic et al., 2018). Similarly, informing individuals that the majority of them act in an environmentally-friendly way can enhance towels reuse in hotels (Goldstein et al., 2008) while placing signs encourage the use of stairs (Brownell et al., 1980; Blamey et al., 1995).
- The consumption of vegetarian meals increases when the menu favors meat-free dishes (Campbell-Arvai et al., 2014). Decreasing the default plate size limits food waste (Kallbekken and Sælen, 2013).
- Offering free bus tickets can reduce drivers' negative perception of public transport and promote persistent bus use (Fujii and Kitamura, 2003; Taniguchi and Fujii, 2007; Beirão and Cabral, 2007).
- Expressing vehicles' fuel efficiency in terms of consumption per distance (e.g. liters per 100 kilometers) instead of distance per consumption (miles per gallon) could help to promote the adoption of efficient vehicles by changing individuals' perception (Larrick and Soll, 2008).
- Showing households the electricity consumption of their neighbors in combination with injunctive norms (e.g. smileys to indicate good performance) enables to decrease the electricity consumption. For example, in the United States, the OPOWER program has reduced the electricity consumption by 2%, which is equivalent to a short-run electricity price increase of 11 to 20% (Allcott, 2011). However, in the absence of injunctive norms there is a risk of a boomerang effect, i.e. the most efficient households actually increase their consumption, another evidence of conformity-driven behavior (Schultz et al.,

[13] Even though nudges do not limit the freedom of choice, individuals could still perceive them as intrusive (Hagman et al., 2015) .

2007). Finally, this effect is heterogeneous and depends on individuals' political views (Costa and Kahn, 2013).

- The share of individuals purchasing green electricity increases when the default electricity mix is the green mix (Pichert and Katsikopoulos, 2008), as illustrated in Section 3.5.1. This effect has also been observed in paper use, energy efficiency and smart grids among others (See Sunstein and Reisch, 2014, for a review of the effects of green defaults).

By contrast with a model assuming *homo oeconomicus* individuals, our framework allows to explain these observations and why nudges are effective. Indeed, nudges rely on individuals' morality and they modify the perceived individual cost, social benefit and cooperation share. They also take advantage of social norms and of the problem framing. Hence, nudges can be a good complement or substitute to financial incentives when the latter fail. However, properly assessing the effects of nudges in our model would require additional information on how nudges influence individuals' perception and the problem framing. Unfortunately, learning about an optimal nudge could be very complex or even impossible (Benkert and Netzer, 2018).

3.6.3 Signaling, learning and education

Instead of relying on financial incentives, policy makers could take advantage of individuals' inclination to contribute voluntarily. One issue hindering individuals' cooperation is the asymmetry of information between producers and consumers. For instance, households have little knowledge about their resource consumption: in Detsouli (2018)'s study, 90% of respondents indicated that they do not know their electricity consumption (see Section 3.5.1). Thus, promoting the implementation of smart meters in combination with feedback on the evolution of households' consumption can help enhance energy conservation (Steg, 2008; Hargreaves et al., 2010; Anda and Temmen, 2014) and curb water demand (Willis et al., 2013; Fielding et al., 2013; Sønderlund et al., 2016). Similarly, knowing the environmental externality associated with the purchase of goods might be challenging. Hence, signaling the environmental quality of products can help trigger behavioral changes. For example, eco-labels and certificates can promote energy-efficiency (Banerjee and Solomon, 2003; Brounen and Kok, 2011) and sustainable food (Caswell and Mojduszka, 1996; Teisl et al., 2002; Brécard et al., 2009). However, an abundance of labels could create confusion. To be effective, labels must also be trusted, easily-understood and accessible (Golan et al., 2001; Banerjee and Solomon, 2003). Else, the cost of implementation might outweigh the benefit.

Another issue hampering environmental-friendly behaviors could be the lack of environmental knowledge in the population. As discussed in Section 3.5.1, a majority of individuals do not know the composition of their electricity mix. Frick et al. (2004) have also shown that Swiss citizens had a poor understanding of greenhouse gas effects, of energy efficiency and of the ozone layer depletion. Increasing the environmental knowledge thanks to education and communication campaigns can support pro-environmental behaviors and lead to better-informed decisions (Bradley et al., 1999; Zsóka et al., 2013). To be effective, the knowledge transmission between experts and the general public requires policy makers to understand what people know. Individuals should also trust the information. In particular, information should be understandable and transparent. The EUCalc project goes in this direction.[14] It aims at developing an online and open-source model in which users can design their own sustainable pathways for European societies by modifying lifestyles and technology development in different sectors (like buildings, transport, agriculture, industry, or energy) thanks to transition levers. The model then computes the environmental (e.g. energy, material and water use, biodiversity) and socio-economic (e.g. health, employment) impacts, and the user can visualize the effects of her scenario in real-time.[15]

As a drawback, environmental signals and education can only influence the behaviors of the most moral individuals. This could explain the observed gap between environmental knowledge and pro-environmental behaviors (Kuhlemeier et al., 1999; Kollmuss and Agyeman, 2002) and why labels tend to widen differences between consumers (Moorman, 1996). In addition, if the most moral individuals also have the highest environmental knowledge, then education would have a limited effect.

Moreover, misperception is not only due to a lack of environmental awareness or knowledge. Beliefs are also shaped by social norms, religion or the emotional context. For instance, the willingness to pay to save birds following the *Exxon Valdez* oil spill was independent of the number of birds that could be saved, individuals reacting to *"the awful image of a helpless bird drowning, its feathers soaked in thick oil"* (Kahneman, 2011, p. 92).[16] In this case, trying to influence beliefs using rational arguments is condemned to failure. Thus, educational campaigns should also rely on strong symbols that reach individuals' affect, such as pictures showing the rapid decline of glaciers or polar bears stuck on ice floe. The question is then how long this effect can last.

Nonetheless, even if education does not directly or permanently change the perceived social benefit, it could still promote cooperative behaviors by influencing the distribution of morality in the population. Bay-Hinitz et al. (1994) have shown that when children play cooperative games instead of competitive games, selfish and aggressive behaviors decreases. This suggests that the educational system should put more emphasis on cooperation between students and less on competition.

3.6.4 Urban planning, infrastructures development and local economy

An appropriate urban planning and the development of green infrastructures can decrease the (perceived) individual cost and thus favor environmental-friendly behaviors. As argued

[14] To learn more about EUCalc: http://www.european-calculator.eu/

[15] I am involved in the EUCalc project, working on the socio-economic impacts and on water management. In particular, I am designing the employment and economic modules and I am supervising the elaboration of the water module.

[16] More precisely, the average contribution to save 2'000, 20'000 and 200'000 birds was \$80, \$78 and \$88 (Kahneman, 2011, p. 92).

in Section 3.5.2, the lack of charging infrastructure is a hurdle to the development of electric vehicles. Similarly, improving public transport availability and connectivity could foster its use. Walking and cycling can be encouraged by providing close-to-home services (e.g. shopping, schools), by creating parks and recreational areas and by building walking and biking trails and secure parking for bicycles (Sallis et al., 1998; Ogilvie et al., 2007).

In addition, the urban planning can also influence the perception of the social benefit. Indeed, public green spaces contribute towards community attachment (Arnberger and Eder, 2012) and pro-environmental attitudes by enhancing the emotional connection with Nature (Budruk et al., 2009). Hence, green cities could support sustainability not only by improving energy efficiency and resource use, but also by affecting individuals' behaviors.

This vision echoes the architecture ideals "Garden-City" of Ebenezer Howard and "Ville Radieuse" of Le Corbusier. These utopian cities include large green spaces nurturing the connection between humans and Nature and promoting biodiversity. They encourage soft mobility by separating streets and walking paths and by locating services and jobs at a walking distance of homes. The "Garden-City" also aims at achieving local self-sufficiency. For example, food production is located inside the city, fostering waste reuse and limiting transport pollution.

Besides decreasing environmental externality, promoting local exchanges and initiatives has other advantages. First, by increasing the assortment between individuals, it could also increase the morality in the population. Second, while economic globalization might give the impression that resources are abundant and conceal environmental pollution, a more local economy could help individuals understand the effects of their consumption and better apprehend the scarcity of resources.

3.7 Lessons learned

In this chapter, we designed a model with heterogeneous moral agents in order to better comprehend individuals' decision making in social dilemmas. Our definitions of morality and of heterogeneous moral population have a strong theoretical foundation, building on the findings of Chapter 2 where we showed that heterogeneous moral populations are favored by evolution. Our framework can explain why individuals care for Nature and why some of them are willing to contribute voluntarily to environmental protection, even though the impact of their actions on environmental externalities is negligible. We then explored the bias affecting individuals' decision-making such as peer pressure, switching costs and misperception. We illustrated how our model can apply to a variety of contexts and can shed light on many empirical findings. Finally, we discussed some policy implications arguing that financial incentives are not always effective and that policy makers could instead rely on other instruments such as nudges, labels and educational campaigns. They could also target different groups with different policies and rely on a combination of instruments. The effectiveness of policies depend on the socio-economic and geographic characteristics of a country. Hence, there does not exist an optimal instrument independently of the context.

Focusing mainly on the effects of morality and its diversity, our settings abstract from other important determinants of individuals' behaviors. In particular, we did not incorporate individuals' wealth such that income and price effects are incorrectly portrayed by the individual cost of cooperating. Since poorer households might not have the financial capacity to invest in environmental-friendly products and technologies (e.g. electric vehicles, rooftop solar panels, insulation), they also do not have the opportunity to behave morally. Introducing income heterogeneity would be crucial to understand the tradeoff between economic and moral drivers and would bring new insights for policy makers.

Furthermore, we did not include other individual motivations such as altruism, empathy, risk-aversion and time-preference. This could hold important consequences for instance for the adoption of technologies and in the context of climate change. Indeed, climate change impacts are future, uncertain and spatially-distributed. Thus, even a fully-moral individual would not make efforts to decrease her greenhouse gas emissions if she has a strong preference for the present. In this case, her perceived social benefit is insignificant. Similarly, since *homo moralis* individuals only consider their own social benefit, they would not care if the impacts affect others. Consequently, morality is only part of the equation. As for policies, economic models and the representation of individuals' preferences should be adjusted to the situation at hand and to empirical evidence.

4 Conclusion

In this thesis we discussed the diversity of social preferences, its theoretical foundations and its implications in the context of environmental, energy and resource economics. First, we explored if heterogeneous populations can be favored by evolution in an evolutionary game theory framework under assortative matching with imperfect information. Introducing the concepts of *evolutionarily stable population* and assortment matrix, we analyzed the evolutionary stability of a heterogeneous population composed of fully-selfish individuals, *homo oeconomicus*, and fully-moral ones, *homo kantiensis*. We showed that there exists a heterogeneous *evolutionarily stable population* for some but but not all games and assortment structures. Consequently, the preferences that are favored by evolution depend on the socio-economic environment and on cultural and geographic conditions. In particular, the assortment structure plays a crucial role by allowing for a greater diversity of *evolutionarily stable populations* and enhancing their robustness to the invasion of mutants. Moreover, both *homo oeconomicus* and *homo kantiensis* are important for the evolutionary success of the population. While *homo kantiensis* individuals drive up the average fitness of the population so that the population is evolutionarily stable for higher values of assortativity, *homo oeconomicus* individuals prevent the invasion of mutants by lowering their fitness.

Building on these findings, we then designed a model with heterogeneous moral agents involved in a social dilemma to shed light on pro-environmental behaviors. Our framework can explain why some individuals are willing to voluntarily engage in costly green actions even though the impact of their efforts on environmental externalities is negligible. We examined how individuals' beliefs can alter their behaviors and hinder cooperation, with an illustration to the purchase of green electricity. Extending our basic model, we showed how incorporating heterogeneous moral agents can provide insights in a variety of contexts such as technology adoption or the purchase of sustainable food. Finally, by better accounting for the social motives behind individuals' decisions, our setting could help policy makers design more effective policies. We notably argue that the effectiveness of policies is shaped by the socio-economic, geographic and regulatory environment of a country. Consequently, financial incentives are not always the most effective instrument and could even fail to promote pro-environmental behaviors. When this is the case, policy makers could instead rely on a combination of instruments such as nudges, labels and educational campaigns.

Furthermore, the design of effective policies requires information on the individuals' characteristics. Thus, an essential step in future research is to estimate the distribution of morality in the population, and its correlation with the income distribution as well as with other motives behind individuals' decisions such as trust, time preferences and attitudes toward risk. Our model should also be tested both in experiments and in real-life contexts (e.g. recycling efforts, the purchase of green electricity and sustainable food and the adoption of technologies) to assess its validity. It could be applied to other public goods such as the contribution to online knowledge and to common goods, adapting the setting to the case of finite populations and relaxing the atomicity assumption when necessary. It would also be interesting to incorporate morality in a macroeconomic model in order to better portray general equilibrium effects. Last but not least, even though we focused on the social motives of individuals, it could be intriguing to examine the objective function of firms in light of an evolutionary framework. Would Milton Friedman (1953, p. 15)'s claim that *"unless the behavior of businessmen in some way or other approximated behavior consistent with the maximization of returns, it seems unlikely that they would remain in business for long"* hold true?

To conclude, this thesis aims at opening the way towards better consideration of the diversity of social preferences, moving away from the classical use of representative agents and homogeneous selfish individuals in economic models. After all, individuals do not live on isolated islands.

Appendix

A The algebra of assortative matching: Proofs

In this section, we provide the proofs of properties, lemmas and proposition of Section 2.2.2 on assortative encounters.

We are in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ (equivalently $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$). Let $I = \{1, 2, \tau\}$, the assortment matrix is $\Phi = ((\phi_{ij}(\lambda, \lambda_\tau)))_{(i,j)\in I^2}$ such that for all $(i, j) \in I^2$, $\phi_{ij}(\lambda, \lambda_\tau) = p_{i|i} - p_{i|j}$ (Definition 1). To be well defined, the matching process must satisfy two sets of conditions:

- The matching conditions: for all $i \in I$, $\sum\limits_{j\in I} p_{j|i} = 1$ (Property 1)
- The balancing conditions: for all $(i, j) \in I^2$, $\lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i}$ (Property 2)

A.1 Proof of Property 3

Property (Assortment balancing condition)**.** The assortment matrix satisfies the *assortment balancing conditions* when:

$$\forall\, (i, j) \in I^2: \quad \lambda_j \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{ik} \right) - \phi_{ij} \right] = \lambda_i \cdot \left[\left(\sum_{k\in I} \lambda_k \phi_{jk} \right) - \phi_{ji} \right]$$

If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions.

Proof.

$$\begin{aligned}
& \lambda_j \cdot \left[\left(\sum_{k\in I}\lambda_k\phi_{ik}\right) - \phi_{ij}\right] - \lambda_i \cdot \left[\left(\sum_{k\in I}\lambda_k\phi_{jk}\right) - \phi_{ji}\right] \\
\underset{\text{(Def.1)}}{=} & \sum_{k\in I}\lambda_j\lambda_k p_{i|i} - \sum_{k\in I}\lambda_j\lambda_k p_{i|k} - \lambda_j p_{i|i} + \lambda_j p_{i|j} - \sum_{k\in I}\lambda_i\lambda_k p_{j|j} + \sum_{k\in I}\lambda_i\lambda_k p_{j|k} + \lambda_i p_{j|j} - \lambda_i p_{j|i} \\
\underset{\text{(Prop.2)}}{=} & \lambda_j p_{i|i} - \sum_{k\in I}\lambda_j\lambda_i p_{k|i} - \lambda_j p_{i|i} + \lambda_i p_{j|i} - \lambda_i p_{j|j} + \sum_{k\in I}\lambda_i\lambda_j p_{k|j} + \lambda_i p_{j|j} - \lambda_i p_{j|i} \\
= & \lambda_i\lambda_j\left[\sum_{k\in I} p_{k|j} - \sum_{k\in I} p_{k|i}\right] \\
\underset{\text{(Prop.1)}}{=} & 0
\end{aligned}$$

□

A.2 Proof of Lemma 1

Lemma (Assortment between residents)**.** *When $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$, if the matching process satisfies the matching and balancing conditions, then we have $\phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0)$.*

Proof. If the matching process satisfies the matching and balancing conditions, then the assortment matrix must satisfy the assortment balancing conditions (Property 3). The assortment balancing conditions are:

$$\begin{aligned}
\lambda_2\left(\lambda_2\phi_{12} + \lambda_\tau\phi_{1\tau} - \phi_{12}\right) &= \lambda_1\left(\lambda_1\phi_{21} + \lambda_\tau\phi_{2\tau} - \phi_{21}\right) \\
\lambda_\tau\left(\lambda_2\phi_{12} + \lambda_\tau\phi_{1\tau} - \phi_{1\tau}\right) &= \lambda_1\left(\lambda_1\phi_{\tau 1} + \lambda_2\phi_{\tau 2} - \phi_{\tau 1}\right) \\
\lambda_\tau\left(\lambda_1\phi_{21} + \lambda_\tau\phi_{2\tau} - \phi_{2\tau}\right) &= \lambda_2\left(\lambda_1\phi_{\tau 1} + \lambda_2\phi_{\tau 2} - \phi_{\tau 2}\right)
\end{aligned}$$

Rewriting the first equation, we get:

$$\phi_{21} = \frac{\lambda_2(1-\lambda_2)\phi_{12} + \lambda_\tau(\lambda_1\phi_{2\tau} - \lambda_2\phi_{1\tau})}{\lambda_1(1-\lambda_1)}$$

Note that for all $(i,j) \in I^2$, $\phi_{ij} = p_{i|i} - p_{i|j}$ is bounded and belongs to $[-1,1]$, and $\lambda_1, \lambda_2 \in (0,1)$. Thus, $\lim_{\lambda_\tau\to 0}\lambda_\tau(\lambda_1\phi_{2\tau} - \lambda_2\phi_{1\tau}) = 0$. Moreover, let $\lambda(\lambda_\tau) \in (0,1)$ be the share of θ_2 with respect to θ_1. We thus have $\lambda_1 = (1-\lambda(\lambda_\tau))(1-\lambda_\tau)$, and $\lambda_2 = \lambda(\lambda_\tau)(1-\lambda_\tau)$. Then noting $\lambda \in (0,1)$ the share of θ_2 with respect to θ_1 when λ_τ goes to zero, we have: $\lim_{\lambda_\tau\to 0}\lambda_2(1-\lambda_2) = \lim_{\lambda_\tau\to 0}\lambda_1(1-\lambda_1) = \lambda(1-\lambda)$. Consequently, $\lim_{\lambda_\tau\to 0}\phi_{12}(\lambda,\lambda_\tau) = \lim_{\lambda_\tau\to 0}\phi_{21}(\lambda,\lambda_\tau)$. □

A.3 Proof of Proposition 1

Proposition (Matching probabilities)**.** *When the assortment matrix Φ satisfies the assortment balancing conditions (Property 3), the system defined by matching conditions (Property 1),* *balancing conditions (Property 2) and assortment matrix conditions (Definition 1) has a unique solution:*

$$\forall (i,j) \in I^2: \quad p_{i|j} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij}$$

Proof. Let (S) be the system of equations defined by matching conditions, balancing conditions and assortment matrix conditions:

$$(S) \begin{cases} \forall\, i \in I, \sum_{j\in I} p_{j|i} = 1 \\ \forall\, (i,j) \in I^2, \lambda_j \cdot p_{i|j} = \lambda_i \cdot p_{j|i} \\ \forall\, (i,j) \in I^2, \phi_{ij} = p_{i|i} - p_{i|j} \end{cases}$$

Suppose there exists matching probabilities $p_{i|j}$ solutions of the system (S). Since $\sum_{k\in I} p_{k|i} = 1$, we have $\sum_{k\in I} \lambda_i p_{k|i} = \lambda_i$ for all $i \in I$. Using the balancing conditions, we get $\lambda_i - \sum_{k\in I} \lambda_k p_{i|k} = 0$. Moreover, since $\sum_{k\in I} \lambda_k = 1$, we have $p_{i|i} = \sum_{k\in I} \lambda_k p_{i|i}$. Adding these two equations, we obtain $p_{i|i} = \lambda_i + \sum_{k\in I} \lambda_k (p_{i|i} - p_{i|k})$ for all $i \in I$, i.e. $p_{i|i} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik}$. Since for all $(i,j) \in I^2$, $p_{i|j} = p_{i|i} - \phi_{ij}$, we get $p_{i|j} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij}$. We have proven that if a solution of (S) exists, then it must be $p_{i|j} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij}$.

We now show that $q_{i|j} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij}$ is solution of (S) using the assortment balancing conditions. First, $q_{i|j}$ satisfies the matching conditions:

$$\begin{aligned} \forall\, j \in I, \sum_{i\in I} q_{i|j} = \sum_{i\in I} \left[\lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij} \right] &= 1 + \sum_{i\in I} \frac{\lambda_i}{\lambda_j} \left[\sum_{k\in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \\ &= 1 + \frac{1}{\lambda_j} \left[\sum_{k\in I} \lambda_k \phi_{jk} - \sum_{i\in I} \lambda_i \phi_{ji} \right] = 1 \end{aligned}$$

Second, $q_{i|j}$ satisfies the balancing conditions:

$$\forall\, (i,j) \in I^2, \lambda_j q_{i|j} - \lambda_i q_{j|i} = \lambda_j \lambda_i + \lambda_j \left[\sum_{k\in I} \lambda_k \phi_{ik} - \phi_{ij} \right] - \lambda_i \lambda_j - \lambda_i \left[\sum_{k\in I} \lambda_k \phi_{jk} - \phi_{ji} \right] = 0$$

Finally, $q_{i|j}$ satisfies the assortment matrix conditions:

$$\forall\, (i,j) \in I^2, q_{i|i} - q_{i|j} = \lambda_i + \sum_{k\in I} \lambda_k \phi_{ik} - \lambda_i - \sum_{k\in I} \lambda_k \phi_{ik} + \phi_{ij} = \phi_{ij}$$

□

A.4 Proof of Lemma 2

Lemma (Conditional probabilities in a population of two residents and one mutant)**.** *When* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, 0)$*, if Proposition 1 is satisfied, then we have:*

$$
\begin{aligned}
p_{1|1} &= (1-\lambda) + \lambda \cdot \phi_{12} \\
p_{1|2} &= (1-\lambda) \cdot (1-\phi_{12}) \\
p_{1|\tau} &= (1-\lambda) \cdot (1-\sigma) - \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{2|1} &= \lambda \cdot (1-\phi_{12}) \\
p_{2|2} &= \lambda + (1-\lambda) \cdot \phi_{12} \\
p_{2|\tau} &= \lambda \cdot (1-\sigma) + \lambda \cdot (1-\lambda) \cdot \Gamma \\
p_{\tau|1} &= 0 \\
p_{\tau|2} &= 0 \\
p_{\tau|\tau} &= \sigma
\end{aligned}
$$

where $\Gamma = \lim_{\lambda_\tau \to 0} \frac{\phi_{\tau 1} - \phi_{\tau 2}}{\lambda_\tau}$.

Proof. If Proposition 1 is satisfied, the conditional probabilities are:

$$
\begin{aligned}
p_{1|1} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} \\
p_{1|2} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} - \phi_{12} \\
p_{1|\tau} &= \lambda_1 + \lambda_2 \cdot \phi_{12} + \lambda_\tau \cdot \phi_{1\tau} - \phi_{1\tau} \\
p_{2|1} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} - \phi_{21} \\
p_{2|2} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} \\
p_{2|\tau} &= \lambda_2 + \lambda_1 \cdot \phi_{21} + \lambda_\tau \cdot \phi_{2\tau} - \phi_{2\tau} \\
p_{\tau|1} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 1} \\
p_{\tau|2} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2} - \phi_{\tau 2} \\
p_{\tau|\tau} &= \lambda_\tau + \lambda_1 \cdot \phi_{\tau 1} + \lambda_2 \cdot \phi_{\tau 2}
\end{aligned}
$$

We can then calculate the limits of the conditional probabilities when the mutant share λ_τ goes to zero. First note that for all $(i, j) \in I^2$, ϕ_{ij} is bounded, and thus $\lim_{\lambda_\tau \to 0} \lambda_\tau \phi_{ij} = 0$. Also, the definition of assortativity implies that: for all $i \in \{1, 2\}$, $\lim_{\lambda_\tau \to 0} \phi_{\tau i} = \sigma$.

Let $\lambda(\lambda_\tau) \in (0, 1)$ be the share of θ_2 with respect to θ_1. We thus have $\lambda_1 = (1 - \lambda(\lambda_\tau))(1 - \lambda_\tau)$, and $\lambda_2 = \lambda(\lambda_\tau)(1 - \lambda_\tau)$. Then noting $\lambda \in (0, 1)$ the share of θ_2 with respect to θ_1 when λ_τ goes to zero, we have: $\lim_{\lambda_\tau \to 0} \lambda_2 = \lambda$ and $\lim_{\lambda_\tau \to 0} \lambda_1 = (1 - \lambda)$.

From Lemma 1, we also have: $\phi_{12}(\lambda, 0) = \phi_{21}(\lambda, 0) \equiv \phi_{12}$.

Finally, we need to compute the limits of $\phi_{1\tau}$ and $\phi_{2\tau}$. We will use the assortment balancing conditions:

$$\begin{aligned}
\lambda_2\left(\lambda_2\phi_{12}+\lambda_\tau\phi_{1\tau}-\phi_{12}\right)&=\lambda_1\left(\lambda_1\phi_{21}+\lambda_\tau\phi_{2\tau}-\phi_{21}\right)\\
\lambda_\tau\left(\lambda_2\phi_{12}+\lambda_\tau\phi_{1\tau}-\phi_{1\tau}\right)&=\lambda_1\left(\lambda_1\phi_{\tau 1}+\lambda_2\phi_{\tau 2}-\phi_{\tau 1}\right)\\
\lambda_\tau\left(\lambda_1\phi_{21}+\lambda_\tau\phi_{2\tau}-\phi_{2\tau}\right)&=\lambda_2\left(\lambda_1\phi_{\tau 1}+\lambda_2\phi_{\tau 2}-\phi_{\tau 2}\right)
\end{aligned}$$

Rewriting the second and third assortment balancing conditions, we get:

$$\begin{aligned}
\phi_{1\tau}&=\frac{\lambda_2}{1-\lambda_\tau}\phi_{12}+\frac{\lambda_1}{1-\lambda_\tau}\frac{(1-\lambda_1)\phi_{\tau 1}-\lambda_2\phi_{\tau 2}}{\lambda_\tau}\\
\phi_{2\tau}&=\frac{\lambda_1}{1-\lambda_\tau}\phi_{21}+\frac{\lambda_2}{1-\lambda_\tau}\frac{(1-\lambda_2)\phi_{\tau 2}-\lambda_1\phi_{\tau 1}}{\lambda_\tau}
\end{aligned}$$

Taking the limit when λ_τ goes to zero:

$$\begin{aligned}
\lim_{\lambda_\tau\to 0}\phi_{1\tau}&=\lambda\phi_{12}+(1-\lambda)\lim_{\lambda_\tau\to 0}\frac{\left[\lambda(\lambda_\tau)+\lambda_\tau-\lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 1}-\left[\lambda(\lambda_\tau)-\lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 2}}{\lambda_\tau}\\
&=\lambda\phi_{12}+(1-\lambda)\lim_{\lambda_\tau\to 0}\left[(1-\lambda(\lambda_\tau))\phi_{\tau 1}+\lambda(\lambda_\tau)\phi_{\tau 2}+\lambda(\lambda_\tau)\frac{\phi_{\tau 1}-\phi_{\tau 2}}{\lambda_\tau}\right]\\
&=\lambda\phi_{12}+(1-\lambda)\sigma+\lambda(1-\lambda)\Gamma\\
\lim_{\lambda_\tau\to 0}\phi_{2\tau}&=(1-\lambda)\phi_{12}+\lambda\lim_{\lambda_\tau\to 0}\frac{\left[1-\lambda(\lambda_\tau)+\lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 2}-\left[1-\lambda(\lambda_\tau)-\lambda_\tau+\lambda(\lambda_\tau)\lambda_\tau\right]\phi_{\tau 1}}{\lambda_\tau}\\
&=(1-\lambda)\phi_{12}+\lambda\lim_{\lambda_\tau\to 0}\left[(1-\lambda(\lambda_\tau))\phi_{\tau 1}+\lambda(\lambda_\tau)\phi_{\tau 2}-(1-\lambda(\lambda_\tau))\frac{\phi_{\tau 1}-\phi_{\tau 2}}{\lambda_\tau}\right]\\
&=(1-\lambda)\phi_{12}+\lambda\sigma-\lambda(1-\lambda)\Gamma
\end{aligned}$$

where $\Gamma=\lim_{\lambda_\tau\to 0}\frac{\phi_{\tau 1}-\phi_{\tau 2}}{\lambda_\tau}$.

Putting it all together, the limits of the conditional probabilities are:

$$\begin{aligned}
p_{1|1}&=(1-\lambda)+\lambda\cdot\phi_{12}\\
p_{1|2}&=(1-\lambda)\cdot(1-\phi_{12})\\
p_{1|\tau}&=(1-\lambda)\cdot(1-\sigma)-\lambda\cdot(1-\lambda)\cdot\Gamma\\
p_{2|1}&=\lambda\cdot(1-\phi_{12})\\
p_{2|2}&=\lambda+(1-\lambda)\cdot\phi_{12}\\
p_{2|\tau}&=\lambda\cdot(1-\sigma)+\lambda\cdot(1-\lambda)\cdot\Gamma\\
p_{\tau|1}&=0\\
p_{\tau|2}&=0\\
p_{\tau|\tau}&=\sigma
\end{aligned}$$

□

B Analysis of evolutionary stability: Proofs

In this section, we provide the proofs related to the analysis of evolutionary stability. We are in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ (equivalently $s = (\theta_1, \theta_2, \theta_\tau, \lambda_1, \lambda_2, \lambda_\tau)$).

B.1 Proof of Lemma 3

Lemma. *$B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.*
If for all $i \in I$ u_{θ_i} are concave in their first arguments, then $B^{NE}(s) \neq \emptyset$.
The correspondence $B^{NE}(\theta_1, \theta_2, \theta_\tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous.

Proof. This proof extends the proof provided by Alger and Weibull (2013) for a population of two types to a population of three types. It follows similar arguments and reasoning.

First, from the definition of a Bayesian Nash equilibrium (Definition 4), we have that, in a population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, $(x_1, x_2, x_\tau) \in X^3$ is a type-homogeneous Bayesian Nash equilibrium if:

$$\forall i \in I : \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in I} p_{j|i} \cdot u_{\theta_i}(x, x_j)$$

With $\lambda_1 = (1-\lambda)(1-\lambda_\tau)$ and $\lambda_2 = \lambda(1-\lambda_\tau)$, we can rewrite the matching probabilities in function of the assortment functions and population shares (Proposition 1). Thus, we get:

$$\forall i \in I : \quad x_i \in \underset{x \in X}{\operatorname{argmax}} \sum_{j \in I} \left(\left[\lambda_j + \sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \cdot u_{\theta_i}(x, x_j) \right)$$

Fixing the population state s, i.e. fixing $(\theta_i)_{i \in I}$ and $(\lambda, \lambda_\tau) \in (0,1) \times [0,1)$, we note for all $i \in I$ $U_{s,i} : X^4 \to \mathbb{R}$ the functions defined by:

$$U_{s,i}(x, x_1, x_2, x_\tau) = \sum_{j \in I} \left(\left[\lambda_j + \sum_{k \in I} \lambda_k \phi_{jk} - \phi_{ji} \right] \cdot u_{\theta_i}(x, x_j) \right)$$

For all $i \in I$, u_{θ_i} is continuous and thus $U_{s,i}$ is also continuous. Since X is compact, then the solution correspondence $\beta_{s,i} : X^3 \rightrightarrows X$ defined by $\beta_{s,i}(x_1, x_2, x_\tau) = \underset{x \in X}{\operatorname{argmax}} U_{s,i}(x, x_1, x_2, x_\tau)$ are non-empty and compact-valued by the Weierstrass's maximum theorem. Hence, the combined correspondence $B_s : X^3 \rightrightarrows X^3$, defined by $B_s(x_1, x_2, x_\tau) = \times_{i \in I} \beta_{s,i}(x_1, x_2, x_\tau)$ is compact valued and, by Berge's maximum theorem, upper hemi-continuous. Hence, B_s has a closed graph and the set of fixed points of B_s, i.e. $B^{NE}(s) = \{(x_i)_{i \in I} : (x_i)_{i \in I} \in B_s((x_i)_{i \in I})\}$, is closed, so that $B^{NE}(s)$ is compact for each $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau) \in \Theta^3 \times (0,1) \times [0,1)$.

Second, since for all $i \in I$, u_{θ_i} is concave in their first arguments then so are $U_{s,i}$.Thus, B_s is convex-valued and has a fixed point by Kakutani's fixed point theorem, i.e. $B^{NE}(s)$ is non-empty.

Third, fixing $(\theta_i)_{i\in I}$, we write for all $i \in I$ $V_{\theta,i} : X^4 \times (0,1) \times [0,1) \to \mathbb{R}$ the functions defined by:

$$V_{\theta,i}(x,x_1,x_2,x_\tau,\lambda,\lambda_\tau) = \sum_{j\in I}\left(\left[\lambda_j + \sum_{k\in I}\lambda_k\phi_{jk} - \phi_{ji}\right]\cdot u_{\theta_i}(x,x_j)\right)$$

Since for all $(i,j) \in I^2$, u_{θ_i} and ϕ_{ij} are continuous, so are $V_{\theta,i}$. Let $V^*_{\theta,i} : X^3 \times (0,1) \times [0,1) \to \mathbb{R}$ the functions defined by $V^*_{\theta,i}(x_1,x_2,x_\tau,\lambda,\lambda_\tau) = \max_{x\in X} V_{\theta,i}(x,x_1,x_2,x_\tau,\lambda,\lambda_\tau)$. By Berge's maximum theorem, $V^*_{\theta,i}$ are continuous. Moreover, by definition of $B^{NE}(s)$, we have, $(x_1,x_2,x_\tau) \in B^{NE}(s)$ if and only if for all $i \in I$:

$$V^*_{\theta,i}(x_1,x_2,x_\tau,\lambda,\lambda_\tau) - V_{\theta,i}(x,x_1,x_2,x_\tau,\lambda,\lambda_\tau) \geq 0 \quad \forall x \in X$$

Let $<\lambda_t>_{t\in\mathbb{N}} \to \lambda^0$ and $<\lambda_{\tau,t}>_{t\in\mathbb{N}} \to \lambda^0_\tau$, and suppose that $(x_{1,t},x_{2,t},x_{\tau,t}) \in B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda_t,\lambda_{\tau,t})$ and for all $i \in I$, $x_{i,t} \to x^0_i$. By continuity of $V_{\theta,i}$ and $V^*_{\theta,i}$, we have for all $i \in I$:

$$V^*_{\theta,i}(x^0_1,x^0_2,x^0_\tau,\lambda^0,\lambda^0_\tau) - V_{\theta,i}(x,x^0_1,x^0_2,x^0_\tau,\lambda^0,\lambda^0_\tau) \geq 0 \quad \forall x \in X$$

This last results proves that $(x^0_1,x^0_2,x^0_\tau) \in B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda^0,\lambda^0_\tau)$ and therefore that the correspondence $B^{NE}(\theta_1,\theta_2,\theta_\tau,\cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is upper hemi-continuous. □

B.2 Proof of Lemma 4

Lemma. *When the population state is $s^\circ = (\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$, if for all $i \in \{1,2\}$, $\Pi_{\theta_i}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ) > \Pi_{\theta_\tau}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ)$ for all $(x^\circ_1,x^\circ_2,x^\circ_\tau) \in B^{NE}(s^\circ)$ then there exists an $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1,x_2,x_\tau,s) > \Pi_{\theta_\tau}(x_1,x_2,x_\tau,s)$ in all Bayesian Nash equilibria (x_1,x_2,x_τ) in all states $s = (\theta_1,\theta_2,\theta_\tau,\lambda,\lambda_\tau)$ with $\lambda_\tau \in (0,\bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. Suppose that in the population state $s^\circ = (\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$, we have for all $i \in \{1,2\}$, $\Pi_{\theta_i}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ) > \Pi_{\theta_\tau}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ)$ for all $(x^\circ_1,x^\circ_2,x^\circ_\tau) \in B^{NE}(s^\circ)$.

For all $i \in I$, the type-fitness Π_{θ_i} are continuous by continuity of the game payoffs and of the assortment functions. Thus, the strict inequalities hold for all $(\hat{x}_1,\hat{x}_2,\hat{x}_\tau)$ in a neighborhood $U \subset X^3 \times (0,1) \times [0,1)$ of $(x_1,x_2,x_\tau,\lambda^\circ,0)$. Using Lemma 3, we know that $B^{NE}(\theta_1,\theta_2,\tau,\cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_{1,t},x_{2,t},x_{\tau,t}) \in B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda_t,\lambda_{\tau,t})$ for all $t \in \mathbb{N}$, $(\lambda_t,\lambda_{\tau,t}) \to (\lambda^\circ,0)$ and $\langle(x_{1,t},x_{2,t},x_{\tau,t})\rangle_{t\in\mathbb{N}}$ converges, then the limit point (x^*_1,x^*_2,x^*_τ) necessarily belongs to $B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \lambda_{\tau,t} < \bar{\varepsilon}$, $|\lambda_t - \lambda^\circ| < \bar{\varepsilon}$ and $(x_{1,t},x_{2,t},x_{\tau,t},\lambda_t,\lambda_{\tau,t}) \in U$, so that for all $i \in I$, $\Pi_{\theta_i}(x_{1,t},x_{2,t},x_{\tau,t},\lambda_t,\lambda_{\tau,t}) > \Pi_{\theta_\tau}(x_{1,t},x_{2,t},x_{\tau,t},\lambda_t,\lambda_{\tau,t})$. □

B.3 Proof of Lemma 5

Lemma. *When the population state is $s^\circ = (\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$, if there exists $i \in \{1,2\}$ such that $\Pi_{\theta_i}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ) < \Pi_{\theta_\tau}(x^\circ_1,x^\circ_2,x^\circ_\tau,s^\circ)$ with $(x^\circ_1,x^\circ_2,x^\circ_\tau) \in B^{NE}(s^\circ)$ a singleton, then there does not* *exist an $\bar{\varepsilon} > 0$ such that for all $i \in \{1,2\}$: $\Pi_{\theta_i}(x_1, x_2, x_\tau, s) > \Pi_{\theta_\tau}(x_1, x_2, x_\tau, s)$ in all Bayesian Nash equilibria (x_1, x_2, x_τ) in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$ with $\lambda_\tau \in (0, \bar{\varepsilon})$ and $|\lambda - \lambda^\circ| < \bar{\varepsilon}$.*

Proof. Suppose that in the population state is $s^\circ = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, there exists $i \in \{1,2\}$ such that $\Pi_{\theta_i}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ) < \Pi_{\theta_\tau}(x_1^\circ, x_2^\circ, x_\tau^\circ, s^\circ)$ with $(x_1^\circ, x_2^\circ, x_\tau^\circ) \in B^{NE}(s^\circ)$ a singleton.

For all $i \in I$, the type-fitness Π_{θ_i} are continuous by continuity of the game payoffs and of the assortment functions. Thus, the strict inequalities hold for all $(\hat{x}_1, \hat{x}_2, \hat{x}_\tau)$ in a neighborhood $U \subset X^3 \times (0,1) \times [0,1)$ of $(x_1, x_2, x_\tau, \lambda^\circ, 0)$. Using Lemma 3, we know that $B^{NE}(\theta_1, \theta_2, \tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. If $(x_{1,t}, x_{2,t}, x_{\tau,t}) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \lambda_{\tau,t})$ for all $t \in \mathbb{N}$, $(\lambda_t, \lambda_{\tau,t}) \to (\lambda^\circ, 0)$ and $\big\langle (x_{1,t}, x_{2,t}, x_{\tau,t}) \big\rangle_{t \in \mathbb{N}}$ converges, then the limit point (x_1^*, x_2^*, x_τ^*) necessarily belongs to $B^{NE}(s^\circ)$. Since by assumption $B^{NE}(s^\circ)$ is a singleton, we have $(x_1^*, x_2^*, x_\tau^*) = (x_1^\circ, x_2^\circ, x_\tau^\circ)$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \lambda_{\tau,t} < \bar{\varepsilon}$, $|\lambda_t - \lambda^\circ| < \bar{\varepsilon}$ and $(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) \in U$, so that $\Pi_{\theta_i}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t}) < \Pi_{\theta_\tau}(x_{1,t}, x_{2,t}, x_{\tau,t}, \lambda_t, \lambda_{\tau,t})$. □

B.4 Proof of Proposition 2 and Corollary 1

Proposition (Type-fitness equality)**.** *In the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness if and only if:*

1. *When $S_\pi = 0$: $Q_\pi = 0$, i.e. $\phi_{12} = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.*
2. *When $S_\pi \neq 0$: $\lambda^\circ = Q_\pi / \big[(1 - \phi_{12})\, S_\pi\big]$.*

Moreover, if homo oeconomicus and homo kantiensis earn the same type fitness, then $\phi_{12} \in (0,1)$.

Corollary (Type-fitness equality under uniformly-constant assortment)**.** *In the population state $s = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ \in (0,1)$, homo oeconomicus (θ_1) and homo kantiensis (θ_2) earn the same type fitness under uniformly-constant assortment if and only if:*

1. *When $S_\pi < 0$: $(\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD}) < \sigma < (\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma) S_\pi]$.*
2. *When $S_\pi = 0$: $\sigma = (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.*
3. *When $S_\pi > 0$: $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$ and $\lambda^\circ = Q_\pi / [(1-\sigma) S_\pi]$.*

Proof. Suppose that *homo oeconomicus* and *homo kantiensis* earn the same type fitness, i.e. $\Pi_{\theta_1}(D, C, s^\circ) = \Pi_{\theta_2}(D, C, s^\circ)$ with:

$$\begin{aligned}
\Pi_{\theta_1}(D, C, s^\circ) &= \big[(1-\lambda^\circ) + \lambda^\circ \cdot \phi_{12}\big] \cdot \pi^{DD} + \big[\lambda^\circ (1 - \phi_{12})\big] \cdot \pi^{DC} \\
\Pi_{\theta_2}(D, C, s^\circ) &= \big[(1-\lambda^\circ)(1 - \phi_{12})\big] \cdot \pi^{CD} + \big[\lambda^\circ + (1-\lambda^\circ)\phi_{12}\big] \cdot \pi^{CC}
\end{aligned}$$

Hence, the type-fitness equality can be rewritten as:

$$\lambda^{\circ}\left(1-\phi_{12}\right)S_{\pi}=Q_{\pi} \tag{B.1}$$

$$\left(1-\lambda^{\circ}\right)\left(1-\phi_{12}\right)S_{\pi}=R_{\pi} \tag{B.2}$$

Where $Q_{\pi} \equiv \pi^{DD}-\pi^{CD}-\phi_{12}(\pi^{CC}-\pi^{CD})$, $R_{\pi} \equiv \pi^{CC}-\pi^{DC}-\phi_{12}(\pi^{DD}-\pi^{DC})$ and $S_{\pi} \equiv \pi^{CC}+\pi^{DD}-\pi^{CD}-\pi^{DC}$.

We first show that $\phi_{12}<1$. Recall that $\phi_{12} \in [-1,1]$ by definition of the assortment (Definition 1). Suppose that $\phi_{12}=1$. This means that *homo oeconomicus* and *homo kantiensis* individuals only meet individuals of their own type. Thus, the type-fitness of *homo oeconomicus* is $\Pi_{\theta_1}(D,C,s^{\circ})=\pi^{DD}$, and the type-fitness of *homo kantiensis* is $\Pi_{\theta_2}(D,C,s^{\circ})=\pi^{CC}$. Since $\pi^{CC}>\pi^{DD}$ by definition of a prisoner's dilemma, *homo kantiensis* earns a strictly greater type-fitness than *homo oeconomicus*, which contradicts our assumption that the two types earn the same fitness. Hence, $\phi_{12}<1$.

We now distinguish two cases: $S_{\pi}=0$ and $S_{\pi}\neq 0$.

When $S_{\pi}=0$, then $Q_{\pi}=0$ (Equation B.1). Thus, $\phi_{12}=(\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})>0$ because $0<\pi^{DD}-\pi^{CD}<\pi^{CC}-\pi^{CD}$ by definition of a prisoner's dilemma ($\pi^{CD}<\pi^{DD}<\pi^{CC}<\pi^{DC}$), and we are in case 1. of the proposition.[1] Under uniformly-constant assortment $\phi_{12}=\sigma$ and we are in case 2. of the corollary.

When $S_{\pi}\neq 0$, we have $\lambda^{\circ}>0$ and $(1-\lambda^{\circ})>0$ since $\lambda^{\circ}\in(0,1)$ by assumption. Moreover, $(1-\phi_{12})>0$ since $\phi_{12}<1$. Thus, $Q_{\pi}\neq 0$, $R_{\pi}\neq 0$ and Q_{π} and R_{π} are of the same sign than S_{π} (Equations B.1 and B.2). Hence, $Q_{\pi}\cdot R_{\pi}>0$ and $\lambda^{\circ}=Q_{\pi}/\left[\left(1-\phi_{12}\right)S_{\pi}\right]$, and we are in case 2. of the proposition. When $S_{\pi}<0$, then $Q_{\pi}<0$ and $R_{\pi}<0$. Thus, $0<(\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})<\phi_{12}$ and $\phi_{12}<(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD})$ by definition of a prisoner's dilemma, which proves $\phi_{12}>0$. Under uniformly-constant assortment $\phi_{12}=\sigma$ and we are in case 1. of the corollary. Similarly, when $S_{\pi}>0$, then $Q_{\pi}>0$ and $R_{\pi}>0$. Thus, $\phi_{12}<(\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$ and $0<(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD})<\phi_{12}$ by definition of a prisoner's dilemma, which proves $\phi_{12}>0$. Under uniformly-constant assortment $\phi_{12}=\sigma$ and we are in case 3. of the corollary.

For the converse, if one of the two cases of the Proposition (or one of the three cases of the Corollary) is true, then Equation (B.1) is satisfied and *homo oeconomicus* and *homo kantiensis* earn the same type fitness. □

B.5 Proof of Lemma 6 and Corollary 2

Lemma (Difference in type fitness between residents and mutant)**.** *Let a population* $s=(\theta_1,\theta_2,\theta_\tau,\lambda^{\circ},0)$*, with* $\lambda^{\circ}\in(0,1)$*, engaged in a prisoners' dilemma such that the residents earn the same type fitness* Π_θ *for* $(x_1,x_2)\in B^{NE}(\theta_1,\theta_2,\lambda^{\circ})$ *with* $x^1\neq x^2$*. Then, the difference in*

[1] Note that we also have $R_{\pi}=0$ (Equation B.2) so that $\phi_{12}=(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD})$. Indeed, since $S_{\pi}=\pi^{CC}+\pi^{DD}-\pi^{CD}-\pi^{DC}=0$, $\pi^{DD}-\pi^{CD}=\pi^{DC}-\pi^{CC}$ and $\pi^{CC}-\pi^{CD}=\pi^{DC}-\pi^{DD}$.

type-fitness between the residents and the mutant for $(x_1, x_2, x_\tau) \in B^{NE}(s)$ *is:*

$$\begin{aligned}\Pi_\theta - \Pi_{\theta_\tau} &= [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot S_\pi \\ &\quad + [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{CC}-\pi^{CD}) + (1-\alpha_2)(\pi^{DC}-\pi^{DD})]\end{aligned}$$

Corollary (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis and* $\lambda^\circ \in (0,1)$*, engaged in a prisoners' dilemma such that the residents earn the same type fitness* Π_θ *for* $(D,C) \in B^{NE}(\theta_1, \theta_2, \lambda^\circ)$*. Then, the difference in type-fitness between the residents and the mutant for* $(D, C, x_\tau) \in B^{NE}(s)$ *is:*

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. Let $(x_1, x_2, x_\tau) \in X^3$ be a Bayesian Nash equilibrium in the population state $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Lemma 2 and noting $\pi_{ij} \equiv \pi(x_i, x_j)$ and $\Pi_{\theta_i} \equiv \Pi_{\theta_i}(x_1, x_2, x_\tau, s)$ for all $(i, j) \in I^2$, we can write the type fitness of each type:

$$\begin{cases}\Pi_{\theta_1} = (1-\lambda^\circ+\lambda^\circ\phi_{12})\cdot\pi_{11} + \lambda^\circ(1-\phi_{12})\cdot\pi_{12} \\ \Pi_{\theta_2} = (1-\lambda^\circ)(1-\phi_{12})\cdot\pi_{21} + [\lambda + (1-\lambda)\phi_{12}]\cdot\pi_{22} \\ \Pi_{\theta_\tau} = [(1-\lambda^\circ)(1-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot\pi_{\tau 1} + [\lambda^\circ(1-\sigma) + \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot\pi_{\tau 2} + \sigma\cdot\pi_{\tau\tau}\end{cases}$$

We know from Property 5 that $(x_1, x_2) \in B^{NE}(s^\circ)$ with $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$. By assumption, θ_1 and θ_2 earns the same type fitness Π_θ in s°. Consequently, they also earn the same type fitness in all Bayesian Nash equilibria in the population state s, i.e. $\Pi_{\theta_1} = \Pi_{\theta_2} \equiv \Pi_\theta$ because in the state s the residents are matched between them, i.e. $\pi_{1\tau}$ and $\pi_{2\tau}$ do not appear in the expression of their type fitness.

In a finite symmetric 2 × 2 fitness games, let A be the matrix of the payoffs in this game, with π^{ij} the payoff when pure strategy i is played against pure strategy j. The payoff obtained by an individual playing strategy x_i when matched with an individual playing x_j is then: $\pi(x_i, x_j) = \pi_{ij} = x_i^\intercal A x_j$. We can rewrite the payoffs in function of the matrix payoff A:

$$\begin{cases}\Pi_{\theta_1} = x_1^\intercal\left[(1-\lambda^\circ)(1-\phi_{12})Ax_1 + \lambda^\circ(1-\phi_{12})Ax_2\right] + \phi_{12}x_1^\intercal Ax_1 \\ \Pi_{\theta_2} = x_2^\intercal\left[(1-\lambda^\circ)(1-\phi_{12})Ax_1 + \lambda^\circ(1-\phi_{12})Ax_2\right] + \phi_{12}x_2^\intercal Ax_2 \\ \Pi_{\theta_\tau} = x_\tau^\intercal\left[((1-\lambda^\circ)(1-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma)Ax_1 + (\lambda^\circ(1-\sigma) + \lambda^\circ(1-\lambda^\circ)\Gamma)Ax_2\right] + \sigma x_\tau^\intercal Ax_\tau\end{cases}$$

Let $\alpha_1, \alpha_2, \alpha_\tau \in [0,1]$ be the probabilities that $\theta_1, \theta_2, \theta_\tau$ individuals attach to the first pure strategy: $x_1 = (\alpha_1, 1-\alpha_1)$, $x_2 = (\alpha_2, 1-\alpha_2)$ and $x_\tau = (\alpha_\tau, 1-\alpha_\tau)$. Since $x_1 \neq x_2$, there exists $\gamma \in \mathbb{R}$ such that $x_\tau = (1-\gamma)x_1 + \gamma x_2$ $(\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2)$.

From type-fitness equality, we know that $\Pi_{\theta_1} = \Pi_{\theta_2} = \Pi_\theta$. Thus, $(1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} = \Pi_\theta$. We can then write the difference between the payoff of the residents and the payoff of the mutants as follows:

$$\begin{aligned}\Pi_\theta - \Pi_\tau &= (1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} - \Pi_\tau\\ &= [(1-\gamma)\phi_{12} - (1-\gamma)^2\sigma - (1-\gamma)(1-\lambda^\circ)(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_1^\intercal A x_1\\ &+ [-\gamma(1-\gamma)\sigma - (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) - (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_1^\intercal A x_2\\ &+ [-\gamma(1-\gamma)\sigma - \gamma(1-\lambda^\circ)(\phi_{12}-\sigma) + \gamma\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_2^\intercal A x_1\\ &+ [\gamma\phi_{12} - \gamma^2\sigma - \gamma\lambda^\circ(\phi_{12}-\sigma) - \gamma\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot x_2^\intercal A x_2\end{aligned}$$

Rearranging, we get:

$$\begin{aligned}\Pi_\theta - \Pi_\tau &= [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot[x_1^\intercal A x_1 - x_1^\intercal A x_2 - x_2^\intercal A x_1 + x_2^\intercal A x_2]\\ &+ [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot[x_2^\intercal A(x_2-x_1)]\end{aligned}$$

We can further develop this expression, using the pure-strategies payoffs:

$$\begin{cases} x_1^\intercal A x_1 = \alpha_1^2\pi^{11} + \alpha_1(1-\alpha_1)(\pi^{21}+\pi^{12}) + (1-\alpha_1)^2\pi^{22}\\ x_1^\intercal A x_2 = \alpha_1\alpha_2\pi^{11} + \alpha_1(1-\alpha_2)\pi^{12} + (1-\alpha_1)\alpha_2\pi^{21} + (1-\alpha_1)(1-\alpha_2)\pi^{22}\\ x_2^\intercal A x_1 = \alpha_1\alpha_2\pi^{11} + \alpha_2(1-\alpha_1)\pi^{12} + (1-\alpha_2)\alpha_1\pi^{21} + (1-\alpha_1)(1-\alpha_2)\pi^{22}\\ x_2^\intercal A x_2 = \alpha_2^2\pi^{11} + \alpha_2(1-\alpha_2)(\pi^{21}+\pi^{12}) + (1-\alpha_2)^2\pi^{22}\end{cases} \tag{B.3}$$

Therefore:

$$\begin{aligned}&x_1^\intercal A x_1 - x_1^\intercal A x_2 - x_2^\intercal A x_1 + x_2^\intercal A x_2 = (\alpha_1-\alpha_2)^2\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right)\\ &x_2^\intercal A(x_2-x_1) = (\alpha_2-\alpha_1)[\alpha_2(\pi^{11}-\pi^{12}) + (1-\alpha_2)(\pi^{21}-\pi^{22})]\end{aligned} \tag{B.4}$$

Consequently, the difference in type fitness when the share of the mutant goes to zero is:

$$\begin{aligned}\Pi_\theta - \Pi_\tau &= [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right)\\ &+ [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{11}-\pi^{12}) + (1-\alpha_2)(\pi^{21}-\pi^{22})]\end{aligned} \tag{B.5}$$

In a prisoners' dilemma, the first pure strategy is cooperate (C) and the second pure strategy is defect (D). Hence, with $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$, we have:

$$\begin{aligned}\Pi_\theta - \Pi_\tau &= [\gamma(1-\gamma)\sigma + (1-\gamma)\lambda^\circ(\phi_{12}-\sigma) + (1-\gamma)\lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)^2\cdot S_\pi\\ &+ [(\gamma-\lambda^\circ)(\phi_{12}-\sigma) - \lambda^\circ(1-\lambda^\circ)\Gamma]\cdot(\alpha_2-\alpha_1)\cdot[\alpha_2(\pi^{CC}-\pi^{CD}) + (1-\alpha_2)(\pi^{DC}-\pi^{DD})]\end{aligned}$$

When the assortment is uniformly constant, $\phi_{12} = \sigma$ and $\Gamma = 0$. Thus, we obtain:

$$\Pi_\theta - \Pi_\tau = \gamma(1-\gamma)\sigma(\alpha_2-\alpha_1)^2 S_\pi$$

Since *homo oeconomicus* always defect $\alpha_1 = 0$, and since *homo kantiensis* always cooperate $\alpha_2 = 1$. Hence, $\gamma = \alpha_\tau$ and:

$$\Pi_\theta - \Pi_\tau = \alpha_\tau(1-\alpha_\tau)\sigma S_\pi$$

□

B.6 Proof of Lemma 7

Lemma (Difference in type fitness between residents and mutant under uniformly-constant assortment)**.** *Under uniformly-constant assortment, let a population* $s = (\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$*, when* θ_1 *is homo oeconomicus,* θ_2 *is homo kantiensis, engaged in a prisoners' dilemma. Then, we have for any* $(D, C, x_\tau) \in B^{NE}(s)$*:*

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau \Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

Proof. Let $(x_1, x_2, x_\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda, \lambda_\tau)$, using Proposition 1 and noting $\lambda_1 = (1-\lambda)(1-\lambda_\tau)$ and $\lambda_2 = \lambda(1-\lambda_\tau)$, we can write the type fitness of each type:

$$\begin{cases} \Pi_{\theta_1} = (\lambda_1(1-\sigma)+\sigma)\cdot x_1^{\mathsf{T}} A x_1 + \lambda_2(1-\sigma)\cdot x_1^{\mathsf{T}} A x_2 + \lambda_\tau(1-\sigma)\cdot x_1^{\mathsf{T}} A x_\tau \\ \Pi_{\theta_2} = \lambda_1(1-\sigma)\cdot x_2^{\mathsf{T}} A x_1 + (\lambda_2(1-\sigma)+\sigma)\cdot x_2^{\mathsf{T}} A x_2 + \lambda_\tau(1-\sigma)\cdot x_2^{\mathsf{T}} A x_\tau \\ \Pi_{\theta_\tau} = \lambda_1(1-\sigma)\cdot x_\tau^{\mathsf{T}} A x_1 + \lambda_2(1-\sigma)\cdot x_\tau^{\mathsf{T}} A x_2 + (\lambda_\tau(1-\sigma)+\sigma)\cdot x_\tau^{\mathsf{T}} A x_\tau \end{cases}$$

Let $x_1 \neq x_2$,and $\alpha_1, \alpha_2, \alpha_\tau \in [0,1]$ be the probabilities that θ_1,θ_2,θ_τ individuals attach to the first pure strategy: $x_1 = (\alpha_1, 1-\alpha_1)$, $x_2 = (\alpha_2, 1-\alpha_2)$ and $x_\tau = (\alpha_\tau, 1-\alpha_\tau)$. Since $x_1 \neq x_2$, there exists $\gamma \in \mathbb{R}$ such that $x_\tau = (1-\gamma)x_1 + \gamma x_2$ ($\alpha_\tau = (1-\gamma)\alpha_1 + \gamma\alpha_2$).

Therefore:

$$\begin{aligned}(1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} = &\left[(1-\gamma)\lambda_1(1-\sigma) + (1-\gamma)\sigma + (1-\gamma)^2\lambda_\tau(1-\sigma)\right]\cdot x_1^{\mathsf{T}} A x_1 \\ &+ \left[(1-\gamma)\lambda_2(1-\sigma) + \gamma(1-\gamma)\lambda_\tau(1-\sigma)\right]\cdot x_1^{\mathsf{T}} A x_2 \\ &+ \left[\gamma\lambda_1(1-\sigma) + \gamma(1-\gamma)\lambda_\tau(1-\sigma)\right]\cdot x_2^{\mathsf{T}} A x_1 \\ &+ \left[\gamma\lambda_2(1-\sigma) + \gamma\sigma + \gamma^2\lambda_\tau(1-\sigma)\right]\cdot x_2^{\mathsf{T}} A x_2\end{aligned}$$

And:

$$\begin{aligned}\Pi_{\theta_\tau} = &\left[(1-\gamma)\lambda_1(1-\sigma) + (1-\gamma)^2\lambda_\tau(1-\sigma) + (1-\gamma)^2\sigma\right]\cdot x_1^{\mathsf{T}} A x_1 \\ &+ \left[(1-\gamma)\lambda_2(1-\sigma) + \gamma(1-\gamma)\lambda_\tau(1-\sigma) + \gamma(1-\gamma)\sigma\right]\cdot x_1^{\mathsf{T}} A x_2 \\ &+ \left[\gamma\lambda_1(1-\sigma) + \gamma(1-\gamma)\lambda_\tau(1-\sigma) + \gamma(1-\gamma)\right]\cdot x_2^{\mathsf{T}} A x_1 \\ &+ \left[\gamma\lambda_2(1-\sigma) + \gamma^2\lambda_\tau(1-\sigma) + \gamma^2\sigma\right]\cdot x_2^{\mathsf{T}} A x_2\end{aligned}$$

Hence:

$$(1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)\left(x_1^\mathsf{T} Ax_1 + x_2^\mathsf{T} Ax_2 - x_1^\mathsf{T} Ax_2 - x_2^\mathsf{T} Ax_1\right)$$

From Equation B.4, we know that: $x_1^\mathsf{T} Ax_1 - x_1^\mathsf{T} Ax_2 - x_2^\mathsf{T} Ax_1 + x_2^\mathsf{T} Ax_2 = (\alpha_2-\alpha_1)^2\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right)$. Consequently:

$$(1-\gamma)\Pi_{\theta_1} + \gamma\Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)(\alpha_2-\alpha_1)^2\left(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21}\right) \tag{B.6}$$

In a prisoners' dilemma, the first pure strategy is cooperate (C) and the second pure strategy is defect (D) and we defined $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$. Moreover, since *homo oeconomicus* always defect $\alpha_1 = 0$, and since *homo kantiensis* always cooperate $\alpha_2 = 1$. Hence, $\gamma = \alpha_\tau$ and we obtain:

$$(1-\alpha_\tau)\Pi_{\theta_1} + \alpha_\tau\Pi_{\theta_2} - \Pi_{\theta_\tau} = \sigma\alpha_\tau(1-\alpha_\tau)S_\pi$$

□

B.7 Proof of Theorem 1

Theorem (Evolutionary stability of a heterogeneous population of *homo oeconomicus* and *homo kantiensis*)**.** *In a prisoners' dilemma under uniformly-constant assortment when* Θ *is rich, there exists a heterogeneous evolutionarily-stable population of homo oeconomicus and homo kantiensis against all types* $\theta_\tau \notin \Theta_{12}$ *if and only if* $S_\pi > 0$ *and* $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$.
Moreover, if $S_\pi > 0$ *and* $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$*, the cooperation share in the evolutionarily stable population satisfies* $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$.

Proof. Suppose that there exists an *evolutionarily stable population* of *homo oeconomicus* and *homo kantiensis* against all types $\theta_\tau \notin \Theta_{12}$. Then, by definition of evolutionary stability (Definition 6), there exists a state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ such that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ. From Corollary 1, we know that there are only three possible cases:

1. When $S_\pi < 0$: $(\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD}) < \sigma < (\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD})$ and $\lambda^\circ = Q_\pi/[(1-\sigma)S_\pi]$.
2. When $S_\pi = 0$: $\sigma = (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$.
3. When $S_\pi > 0$: $(\pi^{DC}-\pi^{CC})/(\pi^{DC}-\pi^{DD}) < \sigma < (\pi^{DD}-\pi^{CD})/(\pi^{CC}-\pi^{CD})$ and $\lambda^\circ = Q_\pi/[(1-\sigma)S_\pi]$.

Let θ_τ a mutant committed to the strategy $\hat{x_\tau} = (1/2; 1/2)$. Such a mutant exists since the type set is rich by assumption. Note also that $\theta_\tau \notin \Theta_{12}$. Then, $(D, C, \hat{x_\tau})$ is a Bayesian Nash equilibrium in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, \lambda_\tau)$ with $\lambda_\tau \in (0,1)$. Using Lemma 7, we have:

$$\frac{\Pi_{\theta_1} + \Pi_{\theta_2}}{2} - \Pi_{\theta_\tau} = \frac{\sigma S_\pi}{4} \tag{B.7}$$

In the three cases satisfying the type-fitness equality, we have $\sigma > 0$ (else *homo oeconomicus* would dominate). Hence, the sign of the left-hand side of Equation B.7 is the same as the sign of S_π. When $S_\pi \leq 0$, we have:

$$\frac{\Pi_{\theta_1} + \Pi_{\theta_2}}{2} \leq \Pi_{\theta_\tau}$$

Hence, θ_τ earns a greater type fitness than the average type-fitness of the residents in all Bayesian Nash equilibria in all states $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, \lambda_\tau)$ with $\lambda_\tau \in (0,1)$. This means that θ_τ earns a greater type fitness than either θ_1 or θ_2 (or both). Thus, the population of *homo oeconomicus* and *homo kantiensis* does not satisfy the second condition for evolutionary stability, which contradicts our initial assumption. Consequently, the only remaining case is $S_\pi > 0$ and then $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$.

Conversely, suppose that $S_\pi > 0$ and $(\pi^{DC} - \pi^{CC})/(\pi^{DC} - \pi^{DD}) < \sigma < (\pi^{DD} - \pi^{CD})/(\pi^{CC} - \pi^{CD})$. Then, from Corollary 1, we know that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ in their only Bayesian Nash equilibrium (D, C) in the population state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi) \in (0,1)$. Let $\theta_\tau \notin \Theta_{12}$ a mutant and $(D, C, x_\tau) \in B^{NE}(s)$ with $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Corollary 2, we can express the difference in type-fitness between the residents and the mutant:

$$\Pi_\theta - \Pi_\tau = \alpha_\tau(1-\alpha_\tau)\sigma S_\pi$$

We have $\sigma > 0$. Moreover, since $\theta_\tau \notin \Theta_{12}$, the mutant does not cooperate or defect, i.e. $\alpha_\tau \in (0,1)$. Thus, $\alpha_\tau(1-\alpha_\tau) > 0$. Finally, $S_\pi > 0$ by assumption. Hence, $\Pi_\theta - \Pi_\tau > 0$. In other words, we have shown that $\Pi_{\theta_1} > \Pi_\tau$ and $\Pi_{\theta_2} > \Pi_\tau$ for any mutant $\theta_\tau \notin \Theta_{12}$ and for any Bayesian Nash equilibrium $(D, C, x_\tau) \in B^{NE}(s)$, with $s = (\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Using Lemma 4, we can conclude that the population of *homo oeconomicus* and *homo kantiensis* in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ with $\lambda^\circ = Q_\pi/((1-\sigma)S_\pi)$ is evolutionarily stable against all types $\theta_\tau \notin \Theta_{12}$. □

B.8 Proof of Lemma 8

Lemma (*Homo hamiltonensis* behavior in prisoners' dilemma)**.** *Let* $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$, $Q_\pi \equiv \pi^{DD} - \sigma\pi^{CC} - (1-\sigma)\pi^{CD}$ *and* $R_\pi \equiv \pi^{CC} - \sigma\pi^{DD} - (1-\sigma)\pi^{DC}$.
When $\sigma = 0$*, homo hamiltonensis is homo oeconomicus and always defects:* $X_\sigma = \{0\}$.
When $\sigma > 0$*,*

1. *If* $S_\pi < 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \textit{if} \quad R_\pi \leq S_\pi \\ \left\{\frac{S_\pi - R_\pi}{(1+\sigma)S_\pi}\right\}, & \textit{if} \quad R_\pi > S_\pi \quad \textit{and} \quad Q_\pi > S_\pi \\ \{1\}, & \textit{if} \quad Q_\pi \leq S_\pi \end{cases}$$

2. *If* $S_\pi = 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \textit{if} \quad R_\pi < S_\pi \\ [0,1], & \textit{if} \quad R_\pi = S_\pi \\ \{1\}, & \textit{if} \quad R_\pi > S_\pi \end{cases}$$

3. *If* $S_\pi > 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \textit{if} \quad R_\pi < 0 \\ \{0,1\}, & \textit{if} \quad Q_\pi, R_\pi \geq 0 \\ \{1\}, & \textit{if} \quad Q_\pi < 0 \end{cases}$$

Proof. When $\sigma = 0$, *homo hamiltonensis* is *homo oeconomicus* and we have shown in Section 2.2.6 that *homo oeconomicus* always defects, i.e. $X_\sigma = \{0\}$.

When $\sigma > 0$, the proof will use a Proposition derived by Alger and Weibull (2013) about the behavior of *homo hamiltonensis* (and more generally *homo moralis*) in 2 × 2 symmetric games. Recall that π^{ij} denotes the payoff when pure strategy i is played against pure strategy j.

Lemma 9 (Proposition 2 of Alger and Weibull, 2013)**.** *Let*

$$\hat{x}(\sigma) = \min\left\{1, \frac{\pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22}}{(1+\sigma)(\pi^{12} + \pi^{21} - \pi^{11} - \pi^{22})}\right\}$$

When $\sigma > 0$*,*

1. *If* $\sigma > 0$ *and* $\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21} < 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \textit{if} \quad \pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} \leq 0 \\ \{\hat{x}(\sigma)\}, & \textit{if} \quad \pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} > 0 \end{cases}$$

2. *If* $\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21} = 0$*, then*

$$X_\sigma = \begin{cases} \{0\}, & \textit{if} \quad \pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} < 0 \\ [0,1], & \textit{if} \quad \pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} = 0 \\ \{1\}, & \textit{if} \quad \pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} > 0 \end{cases}$$

3. *If* $\sigma > 0$ *and* $\pi^{11} + \pi^{22} - \pi^{12} - \pi^{21} > 0$*, then* $X_\sigma \subseteq \{0,1\}$*.*

Let $S_\pi \equiv \pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}$, $Q_\pi \equiv \pi^{22} - \sigma\pi^{11} - (1-\sigma)\pi^{12}$ and $R_\pi \equiv \pi^{11} - \sigma\pi^{22} - (1-\sigma)\pi^{21}$. In a prisoners' dilemma, the first pure strategy is cooperation and the second pure strategy is defection. Thus, we have: $S_\pi \equiv \pi^{CC} + \pi^{DD} - \pi^{CD} - \pi^{DC}$, $Q_\pi \equiv \pi^{DD} - \sigma\pi^{CC} - (1-\sigma)\pi^{CD}$ and $R_\pi \equiv \pi^{CC} - \sigma\pi^{DD} - (1-\sigma)\pi^{DC}$.

Then point 1 ($S_\pi < 0$) and 2 ($S_\pi = 0$) of the Lemma simply rewrites the Proposition 2 of Alger and Weibull (2013) since $\pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} = R_\pi - S_\pi$. Note that when $S_\pi < 0$, $\hat{x}(\sigma) \geq 1$ only if $R_\pi + \sigma S_\pi \geq 0$, i.e. only if $S_\pi - Q_\pi \geq 0$ because $Q_\pi + R_\pi = (1-\sigma)S_\pi$.

For point 3, when $S_\pi > 0$, we know from Proposition 2 of Alger and Weibull (2013) that $X_\sigma \subseteq \{0,1\}$. Suppose that $R_\pi < 0$, then $\pi^{11} < (1-\sigma)\pi^{21} + \sigma\pi^{22}$ and pure strategy 1 (cooperate) is not a *Hamiltonian strategy*, i.e. $\{1\} \notin X_\sigma$ and $X_\sigma = \{0\}$.[2] Similarly, if $Q_\pi < 0$, then $\pi^{22} < (1-\sigma)\pi^{12} + \sigma\pi^{11}$ and pure strategy 2 (defect) is not a *Hamiltonian strategy*, i.e. $X_\sigma = \{1\}$. The last case is when $Q_\pi \geq 0$ and $R_\pi \geq 0$ (because $Q_\pi + R_\pi = (1-\sigma)S_\pi \geq 0$). Let $x = (\alpha_x, 1-\alpha_x) \in X$, we call π_{x1} the payoff when strategy x is played against the first pure strategy, and π_{xx} the payoff when strategy x is played against strategy x. We have:

$$\begin{aligned}
\pi^{11} - (1-\sigma)\pi_{x1} - \sigma\pi_{xx} &= \pi^{11} - (1-\sigma)\left[\alpha_x\pi^{11} + (1-\alpha_x)\pi^{21}\right] \\
&\quad - \sigma\left[\alpha_x^2\pi^{11} + \alpha_x(1-\alpha_x)(\pi^{12}+\pi^{21}) + (1-\alpha_x)^2\pi^{22}\right] \\
&= (1-\alpha_x)\left[\pi^{11} - (1-\sigma)\pi^{21} - \sigma\pi^{22} + \sigma\alpha_x(\pi^{11}+\pi^{22}-\pi^{12}-\pi^{21})\right] \\
&= (1-\alpha_x)\left[R_\pi + \sigma\alpha_x S_\pi\right] \\
&\geq 0
\end{aligned}$$

Hence, the first pure strategy belongs to X_σ. Similarly, we can show that the second pure strategy also belongs to X_σ. Consequently, $X_\sigma = \{0,1\}$.

□

B.9 Proof of Theorem 2

Theorem (Evolutionarily stable population). *In a symmetric* 2×2 *fitness game under uniformly-constant assortment, let* $s° = (\theta_1, \theta_2, \lambda°)$ *be a heterogeneous population with* $\lambda° \in (0,1)$.
The population $s°$ *is evolutionarily stable against all types* $\theta_\tau \notin \Theta_{12}$ *if:*

- *When* $\sigma = 0$*: for all* $(x_1,x_2) \in B^{NE}(s°)$, $x_1 = x_2 \in X_\sigma$ *and* $\beta_\sigma(x_1)$ *is a singleton.*
- *When* $\sigma > 0$*: for all* $(x_1,x_2) \in B^{NE}(s°)$, $(x_1,x_2) \in X_\sigma^2$, $\beta_\sigma(x_1)$ *and* $\beta_\sigma(x_2)$ *are singleton and for all* $(x_1,x_2) \in B^{NE}(s°)$ *such that* $x_1 \neq x_2$, $Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}}) = \lambda°$.

Conversely, if $(x_1,x_2) \in B^{NE}(s°)$ *is a singleton such that* $(x_1,x_2) \notin X_\sigma^2$ *and if* Θ *is rich, then the population is not evolutionarily stable.*

Proof. The proof will use several intermediate results.

[2] Recall that by Definition 8, $x_\sigma \in X$ is a *Halmiltonian strategy* if and only if for all $x \in X$ $\pi(x_\sigma, x_\sigma) \geq (1-\sigma)\pi(x, x_\sigma) + \sigma\pi(x,x)$.

First, for s° to be evolutionarily stable, the residents θ_1 and θ_2 must first earn the same type-fitness in all Bayesian Nash equilibria $(x_1, x_2) \in B^{NE}(s^{\circ})$. Let $Q_{\pi_{1,2}} \equiv \pi_{11} - \pi_{21} - \sigma(\pi_{22} - \pi_{21})$, $R_{\pi_{1,2}} = \pi_{22} - pi_{12} - \sigma(\pi_{11} - \pi_{12})$ and $S_{\pi_{1,2}} \equiv \pi_{11} + \pi_{22} - \pi_{12} - \pi_{21}$. The following Lemma generalizes Proposition 2 about the type-fitness equality between residents:

Lemma 10. *Let $s^{\circ} = (\theta_1, \theta_2, \lambda^{\circ})$ be a heterogeneous population with $\lambda^{\circ} \in (0,1)$, the type-fitness equality is satisfied if and only if:*

1. $Q_{\pi_{1,2}} = S_{\pi_{1,2}} = 0$*, or*
2. $Q_{\pi_{1,2}} R_{\pi_{1,2}} > 0$*, and* $\lambda^{\circ} = Q_{\pi_{1,2}} / ((1-\sigma) S_{\pi_{1,2}})$*, or*
3. $Q_{\pi_{1,2}} = R_{\pi_{1,2}} = 0$*,* $S_{\pi_{1,2}} \neq 0$ *and* $\sigma = 1$*.*

Proof. From the Proof of Proposition 2 (see Appendix B.4), replacing defection D with x_1 and cooperation C with x_2, we can rewrite the type-fitness equality as:

$$\begin{aligned} \lambda^{\circ} (1-\sigma) S_{\pi_{1,2}} &= Q_{\pi_{1,2}} \\ \left(1-\lambda^{\circ}\right) (1-\sigma) S_{\pi_{1,2}} &= R_{\pi} \end{aligned}$$

Suppose the type-fitness equality is satisfied. When $S_{\pi_{1,2}} = 0$, then $Q_{\pi_{1,2}} = R_{\pi_{1,2}} = 0$. When $S_{\pi_{1,2}} \neq 0$, either $\sigma = 1$ and $Q_{\pi_{1,2}} = R_{\pi_{1,2}} = 0$. Else $\sigma \neq 1$ and $Q_{\pi_{1,2}} \neq 0$, $R_{\pi_{1,2}} \neq 0$. Thus $Q_{\pi_{1,2}}$ and $R_{\pi_{1,2}}$ are of the same sign than $S_{\pi_{1,2}}$, and $\lambda^{\circ} = Q_{\pi_{1,2}} / ((1-\sigma) S_{\pi_{1,2}})$. Conversely, if one of the three cases of the Lemma is satisfied, the type-fitness equality is satisfied. □

Lemma 10 tells us that if there exists $(x_1, x_2) \in B^{NE}(s^{\circ})$ such that cases 1, 2 or 3 are not satisfied, then the residents do not earn the same type-fitness and in turn the population is not evolutionarily stable.

Second, we generalize Lemma 7 on the difference in type fitness between residents and mutant under uniformly-constant assortment. Let $S_{\pi} \equiv \pi^{11} + \pi^{22} - \pi^{12} - \pi^{21}$. Note that $S_{\pi_{1,2}} = (\alpha_2 - \alpha_1)^2 S_{\pi}$ (from Equation B.4 in Appendix B.4).

Lemma 11. *Under uniformly-constant assortment, let a population $s = (\theta_1, \theta_2, \theta_{\tau}, \lambda, \lambda_{\tau})$. For any $(x_1, x_2, x_{\tau}) \in B^{NE}(s)$ such that $x_1 \neq x_2$ we have:*

$$(1-\gamma)\Pi_{\theta_1} + \gamma \Pi_{\theta_2} - \Pi_{\theta_{\tau}} = \sigma \gamma (1-\gamma)(\alpha_2 - \alpha_1)^2 S_{\pi}$$

Where $\gamma = (\alpha_{\tau} - \alpha_1)/(\alpha_2 - \alpha_1)$ (i.e. $\alpha_{\tau} = (1-\gamma)\alpha_1 + \gamma \alpha_2$).

Proof. In Appendix B.6 (Equation B.6). □

Note that when Θ is rich, it is always possible to find a mutant θ_{τ} committed to strategy x_{τ} such that $\alpha_1 < \alpha_{\tau} < \alpha_2$ so that $\gamma > 0$ (see Figure B.1). Now suppose $(x_1, x_2) \in B^{NE}(s^{\circ})$ is a singleton such that $x_1 \neq x_2$. Then $(x_1, x_2, x_{\tau}) \in B^{NE}(\theta_1, \theta_2, \theta_{\tau}, \lambda^{\circ}, 0)$ is also a singleton. Using Lemma 3, we know that $B^{NE}(\theta_1, \theta_2, \tau, \cdot) : (0,1) \times [0,1) \rightrightarrows X^3$ is closed-valued and upper hemi-continuous. Thus, if $(x_{1,t}, x_{2,t}, x_\tau) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda_t, \lambda_{\tau,t})$ for all $t \in \mathbb{N}$, $(\lambda_t, \lambda_{\tau,t}) \to (\lambda^\circ, 0)$ and $\langle (x_{1,t}, x_{2,t}, x_\tau) \rangle_{t\in\mathbb{N}}$ converges, then the limit point (x_1^*, x_2^*, x_τ) necessarily belongs to $B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$. Since $B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ is a singleton, we have $(x_1^*, x_2^*, x_\tau^*) = (x_1, x_2, x_\tau)$. Thus, for any given $\bar{\varepsilon} > 0$, there exists a T such that, for all $t > T$, $0 < \lambda_{\tau,t} < \bar{\varepsilon}$, $|\lambda_t - \lambda^\circ| < \bar{\varepsilon}$ and $\alpha_{1,t} < \alpha_\tau < \alpha_{2,t}$. Then, when $\sigma = 0$, $(1-\gamma)\Pi_{\theta_{1,t}} + \gamma\Pi_{\theta_{2,t}} - \Pi_{\theta_\tau} = 0$ and the mutant earns a greater (or equal) type-fitness than at least one of the resident. Consequently, the population is not evolutionarily stable.

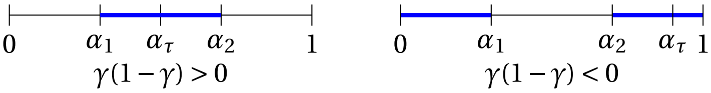


Figure B.1 – Sign of $\gamma(1-\gamma)$ depending on the probabilities attached to the first pure strategy

Third, we will show that if the residents do not play *Hamiltonian strategies*, then the population is not evolutionarily stable.

Lemma 12. *Let $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ be a heterogeneous population with $\lambda^\circ \in (0,1)$. If $(x_1, x_2) \in B^{NE}(s^\circ)$ is a singleton such that $(x_1, x_2) \notin X_\sigma^2$ and if Θ is rich, then the population is not evolutionarily stable.*

Proof. The proof follows two steps. First, we show that there always exists a mutant type that earns a strictly greater type-fitness than the residents at the limit. Then, we extend this result to a small neighborhood.
Let $(x_1, x_2) \in B^{NE}(s^\circ)$ a singleton such that $(x_1, x_2) \notin X_\sigma^2$. Note that if the residents do not earn the same type fitness, the population is not evolutionarily stable. Thus, we will assume next that the residents earn the same type fitness Π_θ.

If $x_1 = x_2 = x_\theta \notin X_\sigma$, then by definition of a *Hamiltonian strategy* (Definition 8), there exists $\hat{x} \in X$ such that $u_\sigma(x_\theta, x_\theta) < u_\sigma(\hat{x}, x_\theta)$, i.e. $\pi(x_\theta, x_\theta) < (1-\sigma)\pi(\hat{x}, x_\theta) + \sigma\pi(\hat{x}, \hat{x})$. At the limit when the population share of the mutant goes to zero, this inequality is equivalent to $\Pi_\theta < \Pi_{\theta_\tau}$, for a mutant playing $\hat{x}$. Moreover, since Θ is rich, there exists a type $\theta_\tau \in \Theta$ for which $\hat{x}$ is strictly dominant, i.e. θ_τ always play $\hat{x}$, and $(x_1, x_2, \hat{x}) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$.

If $x_1 \neq x_2$, from Lemma 11, we know that the difference in type-fitness between the residents and the mutant when the mutant share goes to zero is:

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)(\alpha_2 - \alpha_1)^2 S_\pi$$

We have previously shown that when $\sigma = 0$ the population is not evolutionarily stable. Thus, we turn our attention to the case $\sigma > 0$. We consider the three different cases of Lemma 9:

1. If $S_\pi > 0$, then $X_\sigma \subseteq \{0,1\}$,
   Since Θ is rich, if θ_1 or θ_2 individuals do not play pure strategies, it is always possible to find a mutant θ_τ committed to strategy $\hat{x}$ such that $\gamma(1-\gamma) < 0$ (see Figure B.1). Since

$\sigma > 0$, $\Pi_{\theta} - \Pi_{\theta_{\tau}} < 0$, i.e. the mutant earns a striclty greater type fitness than the residents at the limit in the only Bayesian Nash equilibrium $(x_1, x_2, \hat{x})$.
Else, if θ_1 and θ_2 individuals both play pure strategies, then since $(x_1, x_2) \notin X_{\sigma}^2$, we have $X_{\sigma} = \{0\}$ or $X_{\sigma} = \{1\}$. Thus, one type is playing the Hamiltonian strategy. Without loss of generality and by symmetry, suppose θ_1 individuals are playing the Hamiltonian strategy, and that $X_{\sigma} = \{1\}$ i.e. θ_1 individuals play the first pure strategy while θ_2 individuals play the second pure strategy. We then have $S_{\pi_{1,2}} = S_{\pi} > 0$ and we are in case 2. or 3. of Lemma 10. Thus, we also have $Q_{\pi_{1,2}}, R_{\pi_{1,2}} \geq 0$. Let $x \in X$, such that $x \neq x_2$, i.e. $x = (\eta, 1-\eta)$ with $\eta \in (0,1]$. Then:

$$\begin{aligned}(1-\sigma)\pi(x, x_2) + \sigma\pi(x, x) &= \pi^{22} - \eta R_{\pi_{1,2}} - \sigma\eta(1-\eta)S_{\pi_{1,2}} \\ &\leq \pi^{22}\end{aligned}$$

Thus, for all x in X such that $x \neq x_2$, $u_{\sigma}(x, x_2) \leq u_{\sigma}(x_2, x_2)$. This means that the strategy played by individuals θ_2, i.e. the second pure strategy, is also a Hamiltonian strategy. Consequently, $X_{\sigma} = \{0, 1\}$ which contradicts the assumption $(x_1, x_2) \notin X_{\sigma}^2$. Hence, this case is impossible.

2. If $S_{\pi} = 0$, then we have $S_{\pi_{1,2}} = 0$. Thus, from Lemma 10, we also have $Q_{\pi_{1,2}} = R_{\pi_{1,2}} = 0$. Moreover, using Equation (B.3), we find:

$$Q_{\pi_{1,2}} - S_{\pi_{1,2}} = (\alpha_1 - \alpha_2)[\alpha_2(1+\sigma)S_{\pi} + (\pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22})]$$

Hence, we have $\pi^{12} + \sigma\pi^{21} - (1+\sigma)\pi^{22} = 0$. Therefore, case 2 of Lemma 9 implies that $X_{\sigma} = [0,1]$ which contradicts the assumption $(x_1, x_2) \notin X_{\sigma}^2$, and this case is impossible.

3. If $S_{\pi} < 0$, then since Θ is rich, it is always possible to find a mutant committed to strategy $\hat{x}$ such that $\gamma(1-\gamma) > 0$ (see Figure B.1). Since $\sigma > 0$, $\Pi_{\theta} - \Pi_{\theta_{\tau}} < 0$, i.e. the mutant earns a striclty greater type fitness than the residents at the limit in the only Bayesian Nash equilibrium $(x_1, x_2, \hat{x})$.

Consequently, in the different (possible) cases when $(x_1, x_2) \in B^{NE}(s^{\circ})$ is a singleton such that $(x_1, x_2) \notin X_{\sigma}^2$, we have shown either that the population is not evolutionarily stable or that there exists a mutant type θ_{τ} that earns strictly more than the residents at the limit by being committed to a strategy $\hat{x}$:

$$\begin{aligned} &\Pi_{\theta_1}(x_1, x_2, \hat{x}, \lambda^{\circ}, 0) < \Pi_{\theta_{\tau}}(x_1, x_2, \hat{x}, \lambda^{\circ}, 0) \\ \text{and} \quad &\Pi_{\theta_2}(x_1, x_2, \hat{x}, \lambda^{\circ}, 0) < \Pi_{\theta_{\tau}}(x_1, x_2, \hat{x}, \lambda^{\circ}, 0)\end{aligned}$$

Since $(x_1, x_2, \hat{x}) \in B^{NE}(\theta_1, \theta_2, \theta_{\tau}, \lambda^{\circ}, 0)$, we can conclude using Lemma 5 that the population is not evolutionarily stable. □

Lemma 12 tells us that if the population is evolutionarily stable then the residents play *Hamiltonian strategies.* It also proves the 'Converse' part of the Theorem.

Final intermediate result, when the residents play *Hamiltonian strategies* under uniformly-constant assortment, they earn the same type-fitness:

Lemma 13. *Let $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$ be a heterogeneous population with $\lambda^\circ \in (0,1)$. If $(x_1, x_2) \in B^{NE}(s) \subset X_\sigma^2$ such that $\lambda^\circ = Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}})$ when $x_1 \neq x_2$, then the residents satisfy the type-fitness equality.*

Proof. First, if $x_1 = x_2$, then all individuals play the same strategy and the residents earn the same type fitness. Now suppose that $x_1 \neq x_2$, $(x_1, x_2) \in X_\sigma^2$ and $\lambda^\circ = Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}})$. By definition of a *Hamiltonian strategy* (Definition 8), we have:

$$\begin{cases} x_1 \in \operatorname*{argmax}_{x \in X} u_\sigma(x, x_1) & \Rightarrow \quad \forall x \neq x^1 \in X,\ \pi(x_1, x_1) \geq (1-\sigma) \cdot \pi(x, x_1) + \sigma \cdot \pi(x, x) \\ x_2 \in \operatorname*{argmax}_{y \in X} u_\sigma(y, x_2) & \Rightarrow \quad \forall y \neq x^2 \in X,\ \pi(x_2, x_2) \geq (1-\sigma) \cdot \pi(y, x_2) + \sigma \cdot \pi(y, y) \end{cases}$$

In particular, for $x = x_2$ and $y = x_1$, we have:

$$\begin{cases} \pi_{11} \geq (1-\sigma) \cdot \pi_{21} + \sigma \cdot \pi_{22} & \Rightarrow \quad Q_{\pi_{1,2}} \geq 0 \\ \pi_{22} \geq (1-\sigma) \cdot \pi_{12} + \sigma \cdot \pi_{11} & \Rightarrow \quad R_{\pi_{1,2}} \geq 0 \end{cases}$$

Note that we have $Q_{\pi_{1,2}} + R_{\pi_{1,2}} = (1-\sigma)S_{\pi_{1,2}}$. When $Q_{\pi_{1,2}} = R_{\pi_{1,2}} = 0$, either $S_{\pi_{1,2}} = 0$ and we are in case 1 of Lemma 10 or $\sigma = 1$ and we are in case 3 of Lemma 10. In both cases, the type-fitness equality is satisfied. Now when $Q_{\pi_{1,2}} > 0$ and $R_{\pi_{1,2}} > 0$, since $\lambda^\circ = Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}})$ by assumption, we are in case 2 of Lemma 10 and the residents earn the same type-fitness. Finally, when $Q_{\pi_{1,2}} = 0$ and $R_{\pi_{1,2}} > 0$ then $\lambda^\circ = 0$ and the population is not heterogeneous which contradicts our assumption. Similarly, when $Q_{\pi_{1,2}} > 0$ and $R_{\pi_{1,2}} = 0$, $\lambda^\circ = 1$ and the population is not heterogeneous. □

Let's recap what we have shown for a heterogeneous population in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$

- If there exists $(x_1, x_2) \in B^{NE}(s^\circ)$ such that $x_1 \neq x_2$ and $\lambda^\circ(1-\sigma)S_{\pi_{1,2}} \neq Q_{\pi_{1,2}}$, then then the population is not evolutionarily stable (Lemma 10).
- If $(x_1, x_2) \in B^{NE}(s^\circ)$ is a singleton such that $(x_1, x_2) \notin X_\sigma^2$ and if Θ is rich, then the population is not evolutionarily stable (Lemma 12).
- When $\sigma = 0$, if $(x_1, x_2) \in B^{NE}(s^\circ)$ is a singleton such that $x_1 \neq x_2$ and if Θ is rich, then the population is not evolutionarily stable.

Hence, we still need to show that when the assumptions of the theorem are met, the population is evolutionarily stable.

- When $\sigma > 0$:
  - The population is evolutionarily stable if for all $(x_1, x_2) \in B^{NE}(s^\circ)$, $(x_1, x_2) \in X_\sigma^2$, $\beta_\sigma(x_1)$ and $\beta_\sigma(x_2)$ are singleton and for all $(x_1, x_2) \in B^{NE}(s^\circ)$ such that $x_1 \neq x_2$, $Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}}) = \lambda^\circ$.

- If $(x_1, x_2) \in B^{NE}(s^\circ)$ is a singleton such that $(x_1, x_2) \in X_\sigma^2$ but $\beta_\sigma(x_1)$ or $\beta_\sigma(x_2)$ are not singleton and Θ is rich, then the population is not evolutionarily stable.

When $\sigma = 0$, let $(x_1, x_2) \in B^{NE}(s^\circ)$ such that $x_1 = x_2 = x_\theta \in X_\sigma$. Then, θ_1 and θ_2 earns the same type fitness. Moreover, by definition of a *Hamiltonian strategy* (Definition 8), we have for all $x \in X$, $u_\sigma(x_\theta, x_\theta) \geq u_\sigma(x, x_\theta)$, i.e. $\pi(x_\theta, x_\theta) \geq (1-\sigma)\pi(x, x_\theta) + \sigma\pi(x, x)$. At the limit when the population share of the mutant goes to zero, this inequality is equivalent to $\Pi_\theta \geq \Pi_{\theta_\tau}$, for a mutant playing x. Now if $\beta_\sigma(x_\theta)$ is a singleton, the inequality is strict for all $x \neq x_\theta$. Hence, if for all $(x_1, x_2) \in B^{NE}(s^\circ)$, $x_1 = x_2 \in X_\sigma$ and $\beta_\sigma(x_1)$ is a singleton, we have $\Pi_\theta > \Pi_{\theta_\tau}$ for all Bayesian Nash equilibria (x_1, x_2, x_τ) in the population state $(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ for any mutant $\theta_\tau \notin \Theta_{12}$. Using Lemma 4, we can extend the strict inequality to all Bayesian Nash equilibria in a small neighborhood so that the population is evolutionarily stable.

When $\sigma > 0$, let $(x_1, x_2) \in B^{NE}(s^\circ)$ such that $(x_1, x_2) \in X_\sigma^2$, $\beta_\sigma(x_1)$ and $\beta_\sigma(x_2)$ are singleton and for all $(x_1, x_2) \in B^{NE}(s^\circ)$ such that $x_1 \neq x_2$, $Q_{\pi_{1,2}}/((1-\sigma)S_{\pi_{1,2}}) = \lambda^\circ$.From Lemma 13, we know that θ_1 and θ_2 earns the same type fitness.
If $x_1 = x_2$, following the same arguments as above (when $\sigma = 0$), we can show that $\Pi_\theta > \Pi_{\theta_\tau}$ for all Bayesian Nash equilibria (x_1, x_2, x_τ) in the population state $(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ for any mutant $\theta_\tau \notin \Theta_{12}$.
If $x_1 \neq x_2$, then from Lemma 11, we know that the difference in type-fitness between the residents and any mutant when the mutant share goes to zero is:

$$\Pi_\theta - \Pi_{\theta_\tau} = \sigma\gamma(1-\gamma)(\alpha_2 - \alpha_1)^2 S_\pi$$

Moreover, since $\beta_\sigma(x_1)$ and $\beta_\sigma(x_2)$ are singleton, we have $Q_{\pi_{1,2}} > 0$ and $R_{\pi_{1,2}} > 0$ so that $S_{\pi_{1,2}} > 0$ (because $Q_{\pi_{1,2}} + R_{\pi_{1,2}} = (1-\sigma)S_{\pi_{1,2}}$). Since $S_{\pi_{1,2}} = (\alpha_2 - \alpha_1)^2 S_\pi$, then $S_\pi > 0$ and from Lemma 9 we have $X_\sigma \subseteq \{0, 1\}$. Since $(x_1, x_2) \in X_\sigma^2$ such that $x_1 \neq x_2$, we know that $X_\sigma = \{0, 1\}$. Thus, θ_1 and θ_2 individuals play the two pure strategies. Without loss of generality and by symmetry, we can assume that individuals θ_1 play the pure strategy 2 ($\alpha_1 = 0$), and that individuals θ_2 play the pure strategy 1 ($\alpha_2 = 1$). Thus, γ is the probability that θ_τ attaches to the pure strategy 1, i.e. $\gamma = \alpha_\tau$. When $\theta_\tau \notin \Theta_{12}$, mutants cannot play a pure strategy and $\alpha_\tau \in (0, 1)$ so that $\gamma(1-\gamma) > 0$. We also have $S_\pi > 0$, and $\sigma > 0$. Consequently, the difference in type fitness at the limit is strictly positive, i.e. we have $\Pi_\theta > \Pi_{\theta_\tau}$ for all Bayesian Nash equilibria (x_1, x_2, x_τ) in the population state $(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$ for any mutant $\theta_\tau \notin \Theta_{12}$.
Using Lemma 4, we can extend the strict inequality to all Bayesian Nash equilibria in a small neighborhood so that the population is evolutionarily stable. □

B.10 Proof of Proposition 3

Proposition (Evolutionary stability under state-dependent assortment). *In a prisoners' dilemma, if Θ is rich then there exists $\bar{\sigma} < 1$ such that there does not exist a heterogeneous evolutionary stable population of homo oeconomicus and homo kantiensis for all $\sigma > \bar{\sigma}$.*

Proof. Suppose that *homo oeconomicus* and *homo kantiensis* earn the same type fitness Π_θ in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$. Then, we have $\Pi_\theta = \Pi_{\theta_2} < \pi^{CC}$ because $\Pi_{\theta_2} = p_{1|2} \cdot \pi^{CD} + p_{2|2} \cdot \pi^{CC}$, $\pi^{CD} < \pi^{CC}$ by definition of a prisoners' dilemma and $p_{1|2} > 0$ (since from Proposition 2, $\phi_{12} < 1$). Let $\sigma = 1$ and θ_τ a mutant committed to cooperation. Such a mutant exists since the type set is rich by assumption. Then, the mutants are matched between themselves ($p_{\tau\tau} = 1$) so that $\Pi_{\theta_\tau} = \pi^{CC}$. Hence, we have $\Pi_\theta < \Pi_{\theta_\tau}$ at the limit when the mutant share goes to zero. Since the difference in type fitness between the residents and the mutant is continuous in σ (see Lemma 6), there exists $\bar{\sigma} < 1$ such that the strict inequality holds for all $\sigma > \bar{\sigma}$. Therefore, we have $\Pi_\theta < \Pi_{\theta_\tau}$ for the Bayesian Nash equilibrium $(D, C, C) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, with (D, C, C) a singleton (because each type is committed to its strategy). From Lemma 5, we know that the strict inequality remains valid in a small neighborhood. Consequently, the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable for all $\sigma > \bar{\sigma}$. □

B.11 Proof of Theorem 3

Theorem (Evolutionarily stable population under state-dependent assortment)**.** *Let a population of homo oeconomicus (θ_1) and homo kantiensis (θ_2) in the state $s = (\theta_1, \theta_2, \lambda^\circ)$ involved in a prisoners' dilemma such that the type-fitness equality is satisfied.*
If $(\phi_{12} - \sigma) \notin [\Gamma\lambda^\circ, \Gamma(\lambda^\circ - 1)]$ and if Θ is rich, then the population is not evolutionarily stable. Conversely, when $S_\pi \geq 0$, the population is evolutionarily stable if $(\phi_{12} - \sigma) \in (\Gamma\lambda^\circ, \Gamma(\lambda^\circ - 1))$.

Proof. Let $H : [0,1] \to \mathbb{R}$ be the function that maps the strategy played by the mutant α_τ to the difference in type fitness between the residents and any mutant at the limit when the share of the mutant goes to zero. From Lemma 6, we have for all $\alpha_\tau \in [0,1]$

$$\begin{aligned}H(\alpha_\tau) = &-\alpha_\tau^2 \sigma S_\pi \\ &+ \alpha_\tau \left[\sigma S_\pi - \lambda^\circ(\phi_{12} - \sigma) S_\pi - \lambda^\circ(1-\lambda^\circ)\Gamma S_\pi + (\phi_{12} - \sigma)(\pi^{CC} - \pi^{CD})\right] \\ &+ \lambda^\circ \left[\phi_{12} - \sigma + (1-\lambda^\circ)\Gamma\right](\pi^{DD} - \pi^{DC})\end{aligned}$$

Hence, when the mutants defect or cooperate:

$$\begin{aligned}H(0) &= \lambda^\circ \left[\phi_{12} - \sigma + (1-\lambda^\circ)\Gamma\right](\pi^{DD} - \pi^{DC}) \\ H(1) &= (1-\lambda^\circ)\left[\phi_{12} - \sigma - \lambda^\circ\Gamma\right](\pi^{CC} - \pi^{CD})\end{aligned}$$

Note that we have $\lambda^\circ > 0$, $(1 - \lambda^\circ) > 0$ since the population is assumed heterogeneous, and $(\pi^{CC} - \pi^{CD}) > 0$ and $(\pi^{DD} - \pi^{DC}) < 0$ by definition of a prisoners' dilemma.

Suppose that $(\phi_{12} - \sigma) < \Gamma\lambda^\circ$, then $H(1) < 0$. When Θ is rich, it is always possible to find a mutant θ_τ committed to cooperation. Therefore, we have $\Pi_\theta < \Pi_{\theta_\tau}$ for the Bayesian Nash equilibrium $(D, C, C) \in B^{NE}(\theta_1, \theta_2, \theta_\tau, \lambda^\circ, 0)$, with (D, C, C) a singleton (because each type is committed to its strategy). From Lemma 5, we know that the strict inequality remains valid in a small neighborhood. Consequently, the population of *homo oeconomicus* and *homo* *kantiensis* is not evolutionarily stable.

Similarly, suppose that $(\phi_{12}-\sigma) > \Gamma(\lambda^\circ - 1)$, then $H(0) < 0$. When Θ is rich, it is always possible to find a mutant θ_τ committed to defection. Therefore, we have $\Pi_\theta < \Pi_{\theta_\tau}$ for the Bayesian Nash equilibrium $(D,C,D) \in B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$, with (D,C,D) a singleton. From Lemma 5, we know that the strict inequality remains valid in a small neighborhood. Consequently, the population of *homo oeconomicus* and *homo kantiensis* is not evolutionarily stable.

Conversely, suppose that $(\phi_{12}-\sigma) \in (\Gamma\lambda^\circ, \Gamma(\lambda^\circ - 1))$. Then $H(0) > 0$ and $H(1) > 0$. when $S_\pi \geq 0$, H is concave and attains its minimum on $[0,1]$ in zero or one. Thus, for all $\alpha_\tau \in [0,1]$, we have $H(\alpha_\tau) > 0$. Consequently for any mutant θ_τ, we have $\Pi_\theta > \Pi_{\theta_\tau}$ for any Bayesian Nash equilibrium $(D,C,x_\tau) \in B^{NE}(\theta_1,\theta_2,\theta_\tau,\lambda^\circ,0)$. From Lemma 4, we can conclude that the population of *homo oeconomicus* and *homo kantiensis* is evolutionarily stable. □

B.12 Proof of Proposition 4

Proposition (Robust evolutionarily stable population)**.** *In a prisoners' dilemma under uniformly-constant assortment, there does not exist any robust heterogeneous evolutionarily-stable population of homo oeconomicus and homo kantiensis.*

Proof. Let a a heterogeneous *evolutionarily-stable population* of *homo oeconomicus* and *homo kantiensis* in the state $s^\circ = (\theta_1, \theta_2, \lambda^\circ)$. From Theorem 1, we know that $S_\pi > 0$. Moreover, we know that the difference in type fitness between the two residents is (see e.g. Equation 2.9):

$$\Pi_{\theta_1} - \Pi_{\theta_2} = (1-\lambda^\circ)(1-\sigma)S_\pi - \left[\pi^{CC} - \pi^{DC} - \sigma\left(\pi^{DD} - \pi^{DC}\right)\right]$$

Hence, we have:

$$\frac{\partial\left(\Pi_{\theta_1} - \Pi_{\theta_2}\right)}{\partial\lambda} = -(1-\sigma)S_\pi < 0$$

Consequently, the population is not robust. □

C Cooperation in Social Dilemma: Proofs

In this section, we provide the proofs of Chapter 3 on the cooperation in social dilemmas.

C.1 Proof of Theorem 4

Theorem. *For a given cooperation share of $\bar{x} \in [0,1]$, a homo moralis cooperates if and only if her degree of morality κ_i is greater than the threshold $\kappa_i^0(\bar{x})$ with:*

$$\kappa_i^0(\bar{x}) = \frac{IC_i(\bar{x})}{IC_i(\bar{x}) + SB_i} \quad \in (0,1)$$

Proof. For a given cooperation share of $\bar{x} \in [0,1]$, we know from Equation 3.1 that $u_{\kappa_i}(C,\bar{x}) - u_{\kappa_i}(D,\bar{x}) = -(1-\kappa_i) \cdot IC_i(\bar{x}) + \kappa_i \cdot SB_i$. *Homo moralis* cooperates if and only if $u_{\kappa_i}(C,\bar{x}) - u_{\kappa_i}(D,\bar{x}) \geq 0$, i.e. if and only if $\kappa_i \geq IC_i(\bar{x}/(IC_i(\bar{x}) + SB_i)$ (Recall that by definition of social dilemmas $IC_i(\bar{x} > 0$ and $SB_i > 0)$. □

C.2 Proof of Proposition 5

Proposition. *Let a population of homo moralis involved in a social dilemma such that the degrees of morality are independently drawn from the distribution $F(.)$. There exists an equilibrium cooperation-share $\bar{x}^* \in [0,1]$ such that $\bar{x}^* = 1 - \int_{i \in I} F(\kappa_i^0(\bar{x}^*))d\mu$.*

Proof. Let $G : [0,1] \to \mathbb{R}$ the function such that $G(\bar{x}) = 1 - \int_{i \in I} F(\kappa_i^0(\bar{x}))d\mu - \bar{x}$. $F(.)$ being a CDF, it has values in $[0,1]$. Therefore we have $G(0) = 1 - \int_{i \in I} F(\kappa_i^0(0))d\mu \geq 0$ and $G(1) = -\int_{i \in I} F(\kappa_i^0(1))d\mu \leq 0$. Moreover, for all individuals $i \in I$, $\kappa_i^0 : [0,1] \to (0,1)$ is continuous since individuals' payoffs are continuous in $\bar{x}$. Thus, $F(\kappa_i^0(\cdot))$ is continuous and in turn $\int_{i \in I} F(\kappa_i^0(\cdot))d\mu$ is also continuous. Consequently, $G(\cdot)$ is continuous. Hence, according to the intermediate value theorem, there exists $\bar{x}^* \in [0,1]$ such that $G(\bar{x}^*) = 0$. □

Bibliography

Andrew Abbott, Shasikanta Nandeibam, and Lucy O'Shea. Recycling: Social norms and warm-glow revisited. *Ecological Economics*, 90:10–18, 2013. [46]

Damian C Adams and Matthew J Salois. Local versus organic: A turn in consumer preferences and willingness-to-pay. *Renewable agriculture and food systems*, 25(4):331–341, 2010. [65]

George A Akerlof. Social distance and social decisions. *Econometrica: Journal of the Econometric Society*, pages 1005–1027, 1997. [43, 56]

Ingela Alger and Jörgen W Weibull. A generalization of Hamilton's rule — Love others how much? *Journal of Theoretical Biology*, 299:42–54, 2012. [7]

Ingela Alger and Jörgen W Weibull. Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302, 2013. [1, 2, 3, 4, 5, 6, 11, 13, 15, 28, 37, 49, 82, 91, 92]

Ingela Alger and Jörgen W Weibull. Evolution and Kantian morality. *Games and Economic Behavior*, 98:56–67, 2016. [2, 3, 5, 49]

Ingela Alger and Jörgen W Weibull. Strategic behavior of moralists and altruists. *Games*, 8(3): 38, 2017. [42, 65]

Ingela Alger, Laurent Lehmann, and Jörgen Weibull. Uninvadable social behaviors and preferences in group-structured populations. 2018. [42]

Hunt Allcott. Social norms and energy conservation. *Journal of Public Economics*, 95(9-10): 1082–1095, 2011. [43, 54, 70]

Benjamin Allen and Martin A Nowak. Games among relatives revisited. *Journal of Theoretical Biology*, 378:103–116, 2015. [5, 19, 20]

Francisco Alpizar, Fredrik Carlsson, and Olof Johansson-Stenman. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92(5-6):1047–1060, 2008. [43]

Martin Anda and Justin Temmen. Smart metering for residential energy efficiency: The use of community based social marketing for behavioural change and smart grid introduction. *Renewable Energy*, 67:119–127, 2014. [71]

James Andreoni. Giving with impure altruism: Applications to charity and ricardian equivalence. *Journal of Political Economy*, 97(6):1447–1458, 1989. [46]

James Andreoni. Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):464–477, 1990. [1, 46]

James Andreoni. Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, pages 891–904, 1995. [1]

Arne Arnberger and Renate Eder. The influence of green space on community attachment of urban and suburban residents. *Urban Forestry & Urban Greening*, 11(1):41–49, 2012. [73]

Abhijit Banerjee and Barry D Solomon. Eco-labeling for energy efficiency and sustainability: a meta-evaluation of US programs. *Energy Policy*, 31(2):109–123, 2003. [71]

April K Bay-Hinitz, Robert F Peterson, and H Robert Quilitch. Cooperative games: a way to modify aggressive and cooperative behaviors in young children. *Journal of Applied Behavior Analysis*, 27(3):435, 1994. [72]

Gary S Becker. A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846, 1973. [6]

Gary S Becker. A theory of marriage: Part II. *Journal of Political Economy*, 82(2, Part 2):S11–S26, 1974a. [6]

Gary S Becker. A theory of social interactions. *Journal of Political Economy*, 82(6):1063–1093, 1974b. [1]

Gabriela Beirão and JA Sarsfield Cabral. Understanding attitudes towards public transport and private car: A qualitative study. *Transport Policy*, 14(6):478–489, 2007. [56, 70]

Jean-Michel Benkert and Nick Netzer. Informational requirements of nudging. *Journal of Political Economy*, 126(6):2323–2355, 2018. [71]

Carl T Bergstrom and Peter Godfrey-Smith. On the evolution of behavioral heterogeneity in individuals and populations. *Biology and Philosophy*, 13(2):205–231, 1998. [38]

Theodore C Bergstrom. On the evolution of altruistic ethical rules for siblings. *The American Economic Review*, 85(1):58–81, 1995. [3, 49]

Theodore C Bergstrom. The algebra of assortative encounters and the evolution of cooperation. *International Game Theory Review*, 5(03):211–228, 2003. [5, 7, 9, 11, 19, 20, 23]

Theodore C Bergstrom. Measures of assortativity. *Biological Theory*, 8(2):133–141, 2013. [5]

Helmut Bester and Werner Güth. Is altruism evolutionarily stable? *Journal of Economic Behavior & Organization*, 34(2):193–209, 1998. [5, 12, 39]

Ennio Bilancini, Leonardo Boncinelli, and Jiabin Wu. The interplay of cultural intolerance and action-assortativity for the emergence of cooperation and homophily. *European Economic Review*, 102:1–18, 2018. [5]

Avril Blamey, Nanette Mutrie, and Aitchison Tom. Health promotion by encouraged use of stairs. *BMJ*, 311(7000):289–290, 1995. [70]

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013. [39]

Everett W Bovard Jr. Conformity to social norms and attraction to the group. *Science*, 1953. [43]

Samuel Bowles and Herbert Gintis. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, 65(1):17–28, 2004. [19]

Jennifer Campbell Bradley, Tina M Waliczek, and Jayne M Zajicek. Relationship between environmental knowledge and environmental attitude of high school students. *The Journal of Environmental Education*, 30(3):17–21, 1999. [71]

Charlotte Brannigan, Stephen Luckhurst, Felix Kirsch, Edina Lohr, and Ian Skinner. Ex-post evaluation of directive 2009/33/ec on the promotion of clean and energy efficient road transport vehicles. Technical report, Ricardo Energy & Environment, TEPR, 2018. [62]

Dorothée Brécard, Boubaker Hlaimi, Sterenn Lucas, Yves Perraudeau, and Frédéric Salladarré. Determinants of demand for green products: An application to eco-label demand for fish in europe. *Ecological Economics*, 69(1):115–125, 2009. [71]

Kjell Arne Brekke, Snorre Kverndokk, and Karine Nyborg. An economic model of moral motivation. *Journal of Public Economics*, 87(9-10):1967–1983, 2003. [1, 46]

Kjell Arne Brekke, Karen Evelyn Hauge, Jo Thori Lind, and Karine Nyborg. Playing with the good guys. a public good game with endogenous group formation. *Journal of Public Economics*, 95(9-10):1111–1118, 2011. [1]

Dirk Brounen and Nils Kok. On the economics of energy labels in the housing market. *Journal of Environmental Economics and Management*, 62(2):166–179, 2011. [71]

Kelly D Brownell, Albert J Stunkard, and Janet M Albaum. Evaluation and modification of exercise patterns in the natural environment. *The American Journal of Psychiatry*, 1980. [70]

Annegrete Bruvoll, Bente Halvorsen, and Karine Nyborg. Households' recycling efforts. *Resources, Conservation and recycling*, 36(4):337–354, 2002. [45]

Megha Budruk, Heidi Thomas, and Timothy Tyrrell. Urban green spaces: A study of place attachment and environmental attitudes in India. *Society and Natural Resources*, 22(9): 824–839, 2009. [73]

Hilary Byerly, Andrew Balmford, Paul J Ferraro, Courtney Hammond Wagner, Elizabeth Palchak, Stephen Polasky, Taylor H Ricketts, Aaron J Schwartz, and Brendan Fisher. Nudging pro-environmental behavior: evidence and opportunities. *Frontiers in Ecology and the Environment*, 16(3):159–168, 2018. [70]

Donn Erwin Byrne. *The attraction paradigm*, volume 11. Academic Press, 1971. [7]

Victoria Campbell-Arvai, Joseph Arvai, and Linda Kalof. Motivating sustainable food choices: The role of nudges, value orientation, and information provision. *Environment and Behavior*, 46(4):453–475, 2014. [70]

Stefano Carattini, Andrea Baranzini, Philippe Thalmann, Frédéric Varone, and Frank Vöhringer. Green taxes in a post-paris world: are millions of nays inevitable? *Environmental and Resource Economics*, 68(1):97–128, 2017. [69]

Justin Caron, Thibault Fally, and James R Markusen. International trade puzzles: A solution linking production and preferences. *The Quarterly Journal of Economics*, 129(3):1501–1552, 2014. [67]

Jeffrey P Carpenter. When in rome: conformity and the provision of public goods. *The Journal of Socio-Economics*, 33(4):395–408, 2004. [43]

Eugene M Caruso, Kathleen D Vohs, Brittani Baxter, and Adam Waytz. Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, 142(2):301, 2013. [69]

Julie A Caswell and Eliza M Mojduszka. Using informational labeling to influence the market for quality in food products. *American Journal of Agricultural Economics*, 78(5):1248–1253, 1996. [71]

Rama Cont and Jean-Philipe Bouchaud. Herd behavior and aggregate fluctuations in financial markets. *Macroeconomic Dynamics*, 4(2):170–196, 2000. [43]

Ajla Cosic, Hana Cosic, Sebastian Ille, et al. Can nudges affect students' green behaviour? A field experiment. *Journal of Behavioral Economics for Policy*, 2(1):107–111, 2018. [70]

Dora L Costa and Matthew E Kahn. Energy conservation "nudges" and environmentalist ideology: Evidence from a randomized residential electricity field experiment. *Journal of the European Economic Association*, 11(3):680–702, 2013. [71]

Robin Cowan and Staffan Hultén. Escaping lock-in: the case of the electric vehicle. *Technological forecasting and social change*, 53(1):61–79, 1996. [65]

Sergio Currarini, Matthew O Jackson, and Paolo Pin. An economic model of friendship: Homophily, minorities, and segregation. *Econometrica*, 77(4):1003–1045, 2009. [7, 22, 23]

Robyn M Dawes and Richard H Thaler. Anomalies: cooperation. *Journal of Economic Perspectives*, 2(3):187–197, 1988. [38]

Tiziana De Magistris, Teresa Del Giudice, and Fabio Verneau. The effect of information on willingness to pay for canned tuna fish with different corporate social responsibility (CSR) certification: a pilot study. *Journal of Consumer Affairs*, 49(2):457–471, 2015. [65]

Eddie Dekel, Jeffrey C Ely, and Okan Yilankaya. Evolution of preferences. *The Review of Economic Studies*, 74(3):685–704, 2007. [5, 12, 14, 44]

Mohamed Detsouli. Empirical evidence on non-selfish motives underlying the payment of a premium for green electricity. Technical report, EPFL, 2018. [59, 60, 71]

Ona Egbue and Suzanna Long. Barriers to widespread adoption of electric vehicles: An analysis of consumer attitudes and perceptions. *Energy Policy*, 48:717–729, 2012. [62]

Tore Ellingsen. The evolution of bargaining behavior. *The Quarterly Journal of Economics*, 112 (2):581–602, 1997. [12]

Ilan Eshel and Luigi Luca Cavalli-Sforza. Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, 79(4):1331–1335, 1982. [19]

Carolyn L Evans. The economic significance of national border effects. *The American Economic Review*, 93(4):1291–1312, 2003. [67]

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4): 1645–1692, 2018. [1, 3, 4, 43]

Ricardo Faria, Pedro Moura, Joaquim Delgado, and Anibal T De Almeida. A sustainability assessment of electric vehicles as a personal mobility system. *Energy Conversion and Management*, 61:19–30, 2012. [57, 62]

Ricardo Faria, Pedro Marques, Pedro Moura, Fausto Freire, Joaquim Delgado, and Aníbal T de Almeida. Impact of the electricity mix and use profile in the life-cycle assessment of electric vehicles. *Renewable and Sustainable Energy Reviews*, 24:271–287, 2013. [57, 62]

Ernst Fehr and Simon Gächter. Reciprocity and economics: The economic implications of *Homo Reciprocans*. *European Economic Review*, 42(3-5):845–859, 1998. [1]

Ernst Fehr and Simon Gächter. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181, 2000. [43]

Ernst Fehr and Simon Gächter. Altruistic punishment in humans. *Nature*, 415(6868):137, 2002. [19]

Ernst Fehr and Klaus M Schmidt. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, 1999. [1, 39]

Chaim Fershtman and Yoram Weiss. Social rewards, externalities and stable preferences. *Journal of Public Economics*, 70(1):53–73, 1998. [5]

Kelly S Fielding, Anneliese Spinks, Sally Russell, Rod McCrea, Rodney Stewart, and John Gardner. An experimental test of voluntary strategies to promote urban water demand management. *Journal of Environmental Management*, 114:343–351, 2013. [71]

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, 71(3):397–404, 2001. [1, 54]

Dominique Foray and Arnulf Grübler. Technology and the environment: an overview. *Technological Forecasting and Social Change*, 53(1):3–13, 1996. [65]

Bruno S Frey and Stephan Meier. Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *The American Economic Review*, 94(5): 1717–1722, 2004. [54]

Bruno S Frey and Benno Torgler. Tax morale and conditional cooperation. *Journal of Comparative Economics*, 35(1):136–159, 2007. [59]

Jacqueline Frick, Florian G Kaiser, and Mark Wilson. Environmental knowledge and conservation behavior: Exploring prevalence and structure in a representative sample. *Personality and Individual differences*, 37(8):1597–1613, 2004. [71]

Milton Friedman. *Essays in Positive Economics.* University of Chicago Press, 1953. [76]

Satoshi Fujii and Ryuichi Kitamura. What does a one-month free bus ticket do to habitual drivers? An experimental analysis of habit and attitude change. *Transportation*, 30(1):81–95, 2003. [70]

Luc Gagnon, Camille Belanger, and Yohji Uchiyama. Life-cycle assessment of electricity generation options: The status of research in year 2001. *Energy Policy*, 30(14):1267–1278, 2002. [60]

Uri Gneezy and Aldo Rustichini. A fine is a price. *The Journal of Legal Studies*, 29(1):1–17, 2000. [69]

Elise Golan, Fred Kuchler, Lorraine Mitchell, Cathy Greene, and Amber Jessup. Economics of food labeling. *Journal of Consumer Policy*, 24(2):117–184, 2001. [71]

Noah J Goldstein, Robert B Cialdini, and Vladas Griskevicius. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35(3):472–482, 2008. [43, 54, 70]

Alan Grafen. The hawk-dove game played between relatives. *Animal Behaviour*, 27:905–907, 1979. [39]

Alan Grafen. William Donald Hamilton. 1 august 1936—7 march 2000, 2004. [15]

Anna Gunnthorsdottir, Roumen Vragov, Stefan Seifert, and Kevin McCabe. Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics*, 94 (11-12):987–994, 2010. [6]

Werner Güth and Menahem Yaari. An evolutionary approach to explain reciprocal behavior in a simple strategic game. In U. Witt, editor, *Explaining Process and Change–Approaches to Evolutionary Economics*, pages 23–34. University of Michigan Press, Ann Arbor, 1992. [11]

William Hagman, David Andersson, Daniel Västfjäll, and Gustav Tinghög. Public views on policies involving nudges. *Review of Philosophy and Psychology*, 6(3):439–453, 2015. [70]

William D Hamilton. The genetical evolution of social behaviour. I. *Journal of Theoretical Biology*, 7(1):1–16, 1964a. [20]

William D Hamilton. The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, 7(1):17–52, 1964b. [20]

Andre Hansla, Amelie Gamble, Asgeir Juliusson, and Tommy Gärling. Psychological determinants of attitude towards and willingness to pay for green electricity. *Energy Policy*, 36(2): 768–774, 2008. [59]

Tom Hargreaves, Michael Nye, and Jacquelin Burgess. Making energy visible: A qualitative field study of how householders interact with feedback from smart energy monitors. *Energy Policy*, 38(10):6111–6119, 2010. [71]

Christoph Hauert, Silvia De Monte, Josef Hofbauer, and Karl Sigmund. Volunteering as red queen mechanism for cooperation in public goods games. *Science*, 296(5570):1129–1132, 2002. [19]

Troy R Hawkins, Bhawna Singh, Guillaume Majeau-Bettez, and Anders Hammer Strømman. Comparative environmental life cycle assessment of conventional and electric vehicles. *Journal of Industrial Ecology*, 17(1):53–64, 2013. [57, 62]

Aviad Heifetz, Chris Shannon, and Yossi Spiegel. The dynamic evolution of preferences. *Economic Theory*, 32(2):251–286, 2007. [5, 12]

Martin C Heller and Gregory A Keoleian. Assessing the sustainability of the us food system: a life cycle perspective. *Agricultural Systems*, 76(3):1007–1041, 2003. [65]

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *The American Economic Review*, 91(2):73–78, 2001. [1]

Florian Herold. Carrot or stick? the evolution of reciprocal preferences in a haystack model. *The American Economic Review*, 102(2):914–40, 2012. [5]

William Gord S Hines and John Maynard Smith. Games between relatives. *Journal of Theoretical Biology*, 79(1):19–30, 1979. [39]

Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003. [5]

Herminia Ibarra. Personal networks of women and minorities in management: A conceptual framework. *Academy of Management Review*, 18(1):56–87, 1993. [7]

Ryota Iijima and Yuichiro Kamada. Social distance and network structures. *Theoretical Economics*, 12(2):655–689, 2017. [7]

Matthew O Jackson and Alison Watts. Social games: Matching and the play of finitely repeated games. *Games and Economic Behavior*, 70(1):170–191, 2010. [6]

Martin Kaae Jensen and Alexandros Rigos. Evolutionary games and matching rules. *International Journal of Game Theory*, 47(3):707–735, 2018. [5]

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. [72]

Steffen Kallbekken and Håkon Sælen. Public acceptance for environmental taxes: Self-interest, environmental and distributional concerns. *Energy Policy*, 39(5):2966–2973, 2011. [69]

Steffen Kallbekken and Håkon Sælen. 'nudging' hotel guests to reduce food waste as a win–win environmental measure. *Economics Letters*, 119(3):325–327, 2013. [70]

Immanuel Kant. *Grundlegung zur metaphysik der sitten*, volume 28. L. Heimann, 1870. [1, 49]

Louis Kaplow. Optimal policy with heterogeneous preferences. *The BE Journal of Economic Analysis & Policy*, 8(1), 2008. [1]

Martin G Kocher, Todd Cherry, Stephan Kroll, Robert J Netzer, and Matthias Sutter. Conditional cooperation on three continents. *Economics Letters*, 101(3):175–178, 2008. [54]

Levent Koçkesen, Efe A Ok, and Rajiv Sethi. The strategic advantage of negatively interdependent preferences. *Journal of Economic Theory*, 92(2):274–299, 2000. [5]

Anja Kollmuss and Julian Agyeman. Mind the gap: why do people act environmentally and what are the barriers to pro-environmental behavior? *Environmental Education Research*, 8 (3):239–260, 2002. [72]

Hans Kuhlemeier, Huub Van Den Bergh, and Nijs Lagerweij. Environmental knowledge, attitudes, and behavior in dutch secondary education. *The Journal of Environmental Education*, 30(2):4–14, 1999. [72]

Jean-Jacques Laffont. Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. *Economica*, 42(168):430–437, 1975. [1]

Jessica L Lakin and Tanya L Chartrand. Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, 14(4):334–339, 2003. [7]

Richard P Larrick and Jack B Soll. The MPG illusion. *Science*, 320(5883):1593–1594, 2008. [70]

David F Layton and Gardner Brown. Heterogeneous preferences regarding global climate change. *Review of Economics and Statistics*, 82(4):616–624, 2000. [1]

Robert J Leonard. Reading Cournot, reading Nash: The creation and stabilisation of the Nash equilibrium. *The Economic Journal*, pages 492–511, 1994. [38]

David K Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622, 1998. [1, 56]

Erez Lieberman, Christoph Hauert, and Martin A Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312, 2005. [7]

Gerald Marwell and Ruth E Ames. Economists free ride, does anyone else?: Experiments on the provision of public goods, IV. *Journal of Public Economics*, 15(3):295–310, 1981. [1, 38]

Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y Ozbay. Revealed attention. *The American Economic Review*, 102(5):2183–2205, 2012. [39]

John Maynard Smith. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology*, 47(1):209–221, 1974. [3, 38]

John Maynard Smith and George R Price. The logic of animal conflict. *Nature*, 246(5427): 15–18, 1973. [3, 37]

John McCallum. National borders matter: Canada-us regional trade patterns. *The American Economic Review*, 85(3):615–623, 1995. [67]

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. [7, 22, 54]

Topi Miettinen, Michael Kosfeld, Ernst Fehr, and Jorgen W Weibull. Revealed preferences in a sequential prisoners' dilemma: A horse-race between five utility functions. 2017. [42, 49]

Wanki Moon, Wojciech J Florkowski, Bernhard Brückner, and Ilona Schonhof. Willingness to pay for environmental practices: implications for eco-labeling. *Land Economics*, 78(1): 88–102, 2002. [45]

Christine Moorman. A quasi experiment to assess the consumer and informational determinants of nutrition information processing activities: The case of the nutrition labeling and education act. *Journal of Public Policy & Marketing*, 15(1):28–44, 1996. [72]

John Nash. *Non-cooperative games*. PhD thesis, Princeton, 1950. [38]

John Nash. Non-cooperative games. *Annals of Mathematics*, pages 286–295, 1951. [38]

Nick Netzer. Evolution of time preferences and attitudes toward risk. *The American Economic Review*, 99(3):937–55, 2009. [39]

Jonathan Newton. The preferences of Homo Moralis are unstable under evolving assortativity. *International Journal of Game Theory*, 46(2):583–589, 2017. [42]

Bryan Norton, Robert Costanza, and Richard C Bishop. The evolution of preferences: why sovereign' preferences may not lead to sustainable policies and what to do about it. *Ecological Economics*, 24(2-3):193–211, 1998. [14]

Martin A Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006. [5]

Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291, 2005. [19]

Martin A Nowak, Corina E Tarnita, and Tibor Antal. Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537): 19–30, 2010. [5, 7]

Karine Nyborg, Richard B Howarth, and Kjell Arne Brekke. Green consumers and public policy: On socially contingent moral motivation. *Resource and Energy Economics*, 28(4):351–366, 2006. [46]

Peter Ockenfels. Cooperation in prisoners' dilemma: An evolutionary approach. *European Journal of Political Economy*, 9(4):567–579, 1993. [12]

OECD. Environment at a Glance 2015: OECD Indicators, 2015. [46]

David Ogilvie, Charles E Foster, Helen Rothnie, Nick Cavill, Val Hamilton, Claire F Fitzsimons, and Nanette Mutrie. Interventions to promote walking: systematic review. *BMJ*, 334(7605): 1204, 2007. [73]

Hisashi Ohtsuki and Martin A Nowak. Evolutionary stability on graphs. *Journal of Theoretical Biology*, 251(4):698–707, 2008. [7]

Efe A Ok and Fernando Vega-Redondo. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, 97(2):231–254, 2001. [5, 12]

Daniel Pichert and Konstantinos V Katsikopoulos. Green defaults: Information presentation and pro-environmental behaviour. *Journal of Environmental Psychology*, 28(1):63–73, 2008. [71]

Thomas Piketty. Social mobility and redistributive politics. *The Quarterly Journal of Economics*, 110(3):551–584, 1995. [1]

Robert S Pindyck. Climate change policy: what do the models tell us? *Journal of Economic Literature*, 51(3):860–72, 2013. [58]

Alex Possajennikov. On the evolutionary stability of altruistic and spiteful preferences. *Journal of Economic Behavior & Organization*, 42(1):125–129, 2000. [12]

Matthew Rabin. Incorporating fairness into game theory and economics. *The American Economic Review*, pages 1281–1302, 1993. [1]

Lucia A Reisch and Cass R Sunstein. Do Europeans like nudges? *Judgment and Decision Making*, 11(4):310–325, 2016. [70]

Jörg Rieskamp, Jerome R Busemeyer, and Barbara A Mellers. Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, 44(3): 631–661, 2006. [39]

Arthur J Robson. Efficiency in evolutionary games: Darwin, Nash and the secret handshake. *Journal of Theoretical Biology*, 144(3):379–396, 1990. [12]

Brian Roe, Mario F Teisl, Alan Levy, and Matthew Russell. US consumers' willingness to pay for green electricity. *Energy Policy*, 29(11):917–925, 2001. [59]

Poritosh Roy, Daisuke Nei, Takahiro Orikasa, Qingyi Xu, Hiroshi Okadome, Nobutaka Nakamura, and Takeo Shiina. A review of life cycle assessment (LCA) on some food products. *Journal of Food Engineering*, 90(1):1–10, 2009. [65]

James F Sallis, Adrian Bauman, and Michael Pratt. Environmental and policy interventions to promote physical activity. *American Journal of Preventive Medicine*, 15(4):379–397, 1998. [73]

Catherine Salmon and Margo Wilson. Kinship: The conceptual hole in psychological studies of social cognition and close relationships. *Evolutionary Social Psychology*, page 265, 2013. [7]

William H Sandholm. *Population games and evolutionary dynamics*. MIT press, 2010. [5]

Hannah Schildberg-Hörisch. Are Risk Preferences Stable? *Journal of Economic Perspectives*, 32(2):135–54, 2018. [39]

Bodo B Schlegelmilch, Greg M Bohlen, and Adamantios Diamantopoulos. The link between green purchasing decisions and measures of environmental consciousness. *European Journal of Marketing*, 30(5):35–55, 1996. [1]

P Wesley Schultz, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434, 2007. [70]

Reinhard Selten and Jose Apesteguia. Experimentally observed imitation and cooperation in price competition on the circle. *Games and Economic Behavior*, 51(1):171–192, 2005. [43]

Rajiv Sethi and Eswaran Somanathan. Preference evolution and reciprocity. *Journal of Economic Theory*, 97(2):273–297, 2001. [12]

SFOE. Statistique suisse de l'électricité 2017. Technical report, Swiss Federal Office of Energy SFOE, 2018. [60]

Paulo Shakarian, Patrick Roos, and Anthony Johnson. A review of evolutionary graph theory with applications to game theory. *Biosystems*, 107(2):66–80, 2012. [7]

Jen Shang and Rachel Croson. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *The Economic Journal*, 119 (540):1422–1439, 2009. [43]

Jason F Shogren and Laura O Taylor. On behavioral-environmental economics. *Review of Environmental Economics and Policy*, 2(1):26–44, 2008. [39]

William Sierzchula, Sjoerd Bakker, Kees Maat, and Bert Van Wee. The influence of financial incentives and other socio-economic factors on electric vehicle adoption. *Energy Policy*, 68: 183–194, 2014. [62, 63]

Kristin N Sipes and Robert Mendelsohn. The effectiveness of gasoline taxation to manage air pollution. *Ecological Economics*, 36(2):299–309, 2001. [69]

Adam Smith. *The Theory of Moral Sentiments*. A. Millar; and A. Kincaid and J. Bell, in Edinburgh, 1759. [1]

Anders L Sønderlund, Joanne R Smith, Christopher J Hutton, Zoran Kapelan, and Dragan Savic. Effectiveness of smart meter-based consumption feedback in curbing household water use: Knowns and unknowns. *Journal of Water Resources Planning and Management*, 142(12): 04016060, 2016. [71]

Oded Stark and Ita Falk. Transfers, empathy formation, and reverse transfers. *The American Economic Review*, 88(2):271–276, 1998. [1]

Linda Steg. Promoting household energy conservation. *Energy Policy*, 36(12):4449–4453, 2008. [71]

Swantje Sundt and Katrin Rehdanz. Consumers' willingness to pay for green electricity: A meta-analysis of the literature. *Energy Economics*, 51:1–8, 2015. [45, 59]

Cass R Sunstein and Lucia A Reisch. Automatically green: Behavioral economics and environmental protection. *Harvard Environmental Law Review*, 38:127, 2014. [71]

Ayako Taniguchi and Satoshi Fujii. Promoting public transport using marketing techniques in mobility management and verifying their quantitative effects. *Transportation*, 34(1):37, 2007. [70]

Corina E Tarnita, Tibor Antal, Hisashi Ohtsuki, and Martin A Nowak. Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences*, 106(21): 8601–8604, 2009. [7]

Mario F Teisl, Brian Roe, and Robert L Hicks. Can eco-labels tune a market? Evidence from dolphin-safe labeling. *Journal of Environmental Economics and Management*, 43(3):339–359, 2002. [71]

Richard H Thaler and Cass R Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, 2008. [70]

Philippe Thalmann. The public acceptance of green taxes: 2 million voters express their opinion. *Public Choice*, 119(1-2):179–217, 2004. [69]

Dawn Thilmany, Craig A Bond, and Jennifer K Bond. Going local: Exploring consumer behavior and motivations for direct food purchases. *American Journal of Agricultural Economics*, 90 (5):1303–1309, 2008. [65]

Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management Science*, 39(10):1179–1189, 1993. [39]

Eric Van Damme. Evolutionary game theory. In *Stability and Perfection of Nash Equilibria*, pages 214–258. Springer, 1991. [14]

Edwin JC Van Leeuwen, Katherine A Cronin, Daniel BM Haun, Roger Mundry, and Mark D Bodamer. Neighbouring chimpanzee communities show different preferences in social grooming behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, 279 (1746):4362–4367, 2012. [3]

Riccardo Vecchio. Determinants of willingness-to-pay for sustainable wine: Evidence from experimental auctions. *Wine Economics and Policy*, 2(2):85–92, 2013. [65]

Iris Vermeir and Wim Verbeke. Sustainable food consumption: Exploring the consumer "attitude–behavioral intention" gap. *Journal of Agricultural and Environmental Ethics*, 19(2): 169–194, 2006. [43, 54, 65]

Kathleen D Vohs, Nicole L Mead, and Miranda R Goode. The psychological consequences of money. *Science*, 314(5802):1154–1156, 2006. [69]

Jörgen W Weibull. The mass-action interpretation of Nash equilibrium, 1994. [38]

Andrew Whiten, Victoria Horner, and Frans BM De Waal. Conformity to cultural norms of tool use in chimpanzees. *Nature*, 437(7059):737, 2005. [43]

Rachelle M Willis, Rodney A Stewart, Damien P Giurco, Mohammad Reza Talebpour, and Alireza Mousavinejad. End use water consumption in households: impact of socio-demographic factors and efficient devices. *Journal of Cleaner Production*, 60:107–115, 2013. [71]

Holger C Wolf. Intranational home bias in trade. *Review of Economics and Statistics*, 82(4): 555–563, 2000. [67]

Sewall Wright. Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645): 330–338, 1922. [7]

Kei-Mu Yi. Can multistage production explain the home bias in trade? *The American Economic Review*, 100(1):364–93, 2010. [67]

Min Zhou. Intensification of geo-cultural homophily in global trade: Evidence from the gravity model. *Social Science Research*, 40(1):193–209, 2011. [54]

Ágnes Zsóka, Zsuzsanna Marjainé Szerényi, Anna Széchy, and Tamás Kocsis. Greening due to environmental education? Environmental knowledge, attitudes, consumer behavior and everyday pro-environmental activities of Hungarian high school and university students. *Journal of Cleaner Production*, 48:126–138, 2013. [71]

Curriculum Vitae

BORIS THURM

Contact information

EPFL ENAC IA LEURE	boris.thurm@epfl.ch; +41 21 69 36268
BP 2138 (Bâtiment BP)	https://people.epfl.ch/boris.thurm
Station 16	25.03.1989, French citizen
CH-1015 Lausanne	

Research interests

Environmental, resource and energy economics, (evolutionary) game theory, microeconomic theory, public economics, behavioral economics

Education

PhD *2019*
Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
Dissertation: *The diversity of moral preferences: Evolutionary foundations and some implications in environmental economics*

Swiss Program for Beginning Doctoral Students in Economics *2017*
Study Center Gerzensee, Switzerland

MSc in Energy, Management and Sustainability *2014*
EPFL, Lausanne, Switzerland
EPFL Excellence Fellowship
Master Thesis: *Exploring the possibility of an Integrated Resource Management for UBC - Focus on the Water-Energy Nexus*, written at the University of British Columbia (UBC), Vancouver, Canada

BSc in Environmental Sciences and Engineering *2011*
EPFL, Lausanne, Switzerland

Engineering school preparatory class *Sept. 2007 - June 2009*
Lycée du Parc, Lyon, France

ACADEMIC AND PROFESSIONAL EXPERIENCE

Laboratory of Environmental and Urban Economics (LEUrE) *Jan. 2015 - Present*
EPFL, Lausanne, Switzerland
Doctoral Assistant

- Dissertation under the supervision of Prof. Philippe Thalmann
- Project *EU Calculator: Trade-offs and Pathways towards Sustainable and Low-carbon European Societies*, EUCalc, funded by the European Union's Horizon 2020 research and innovation programme. Working on the socio-economic impacts of decarbonizing European societies, focusing on employment, and on water management. Part of the Management Board since November 2018, representing EPFL
- Project *CCImpact* on the economic impacts of climate change in Switzerland, funded by the Swiss Federal Office for the Environment. Worked on the economic impacts of climate change in the tourism sector
- Supervision of master projects: *Impact of the energy transition on employment in Europe* (2017, 2018) by Lucas Spierenburg, *Environmental and socio-economic impacts of mobility policies in Europe* (2019) by Jean-Baptiste Decoppet and Gauthier de Dreuille
- Teaching assistant: Mise à niveau mathématiques (Mathematics, 1^{st} year EPFL Bachelor), Croissance et développement durable (Growth and sustainable development)

Institute for Resources, Environment and Sustainability (IRES) *April 2014 - Sept. 2014*
University of British Columbia (UBC), Vancouver, Canada

- Master Thesis supervised by Prof. Gunilla Öberg (UBC) and Prof. Matthias Finger (EPFL)
- Project *Would it make sense to develop an integrated resource management strategy for UBC, using a water lens?*

Chair Management of Network Industries (MIR) *Sept. 2012 - June 2013*
Innovative Governance of Large Urban Systems (IGLUS), EPFL, Lausanne, Switzerland
Semester projects: *Toward a sustainable use of water in urban areas - Comparison of San Francisco and Detroit* and *Toward a sustainable use of water in urban areas - Energy and water relationship*, under the supervision of Prof. Matthias Finger and Dr. Mohamad Razaghi (EPFL)

Amaudruz Energies, Lausanne, Switzerland *Oct. 2011 - March 2012*
Internship: design of renewable (thermal solar and PV) and electrical installations, energy balances, thermography analysis, market analysis

La Crêmerie du Moulin Restaurant, Chamonix, France *July-Aug. 2011 and 2012*
Cook, dishwasher

Chamonix Town Hall, France
Public toilet maintenance *June-July 2009 and July-Aug. 2010*
Green-space maintenance, gardener *July 2008*

Intermarche supermarket, Sallanches, France *July 2006*
Shelves filling

RESEARCH

PEER-REVIEWED PUBLICATIONS

- Frank Vöhringer, Marc Vielle, Philippe Thalmann, Anita Frehner, Wolfgang Knoke, Dario Stocker, Boris Thurm (2019). Costs and benefits of climate change in Switzerland. *Climate Change Economics*

WORKING PAPERS AND CONFERENCES

- Charles Ayoubi and Boris Thurm. Why do some individuals care for Nature? Morality and Social Dilemmas. Presented at the 1st *Gerzensee 2016 Alumni Conference* (talk), Gerschnialp, Switzerland, April 13-14, 2018
- Charles Ayoubi and Boris Thurm. Exploring the diversity of social preferences: Is a heterogeneous population evolutionarily stable under assortative matching? Presented at the *AEA/ASSA 2019 Conference* (poster), Atlanta, Georgia, January 4-6, 2019 and at the *Gerzensee Alumni Conference 2017* (talk), Study Center Gerzensee, December 5, 2017. Available here
- Charles Ayoubi and Boris Thurm. The Algebra of Assortative Matching
- Boris Thurm and Marc Vielle. Employment impacts of decarbonizing European societies. Presented at the *Green Jobs Assessment Institutions Network (GAIN), 3rd International Conference: Just Transition* (talk), Geneva, December 6-7, 2017. Available here
- Boris Thurm, Marc Vielle and Frank Vöhringer. Impacts of climate change for Swiss winter and summer tourism: a general equilibrium analysis. Presented at the *EAERE 23rd Annual Conference* (poster), Athens, June 28-30, 2017 and at the *SSES Annual Congress 2017* (talk), Lausanne, June 8-9, 2017. Available here

PROJECT REPORTS

- Gino Baudry, Francesco Clora, Onesmus Mwabonje, Boris Thurm, Jeremy Woods, Wusheng Yu (2018). Deliverable 7.2: Documentation of GTAP-EUCalc interface and design of GTAP scenarios. Public deliverable of the EUCalc Project. Available here
- Farahnaz Pashaei Kamali, Boris Thurm, Ana Rankovic, Marc Vielle , John Posada, Patricia Osseweijer (2018). Deliverable 6.3: Expert consultation workshop on identification of key socio-economic parameters. Public deliverable of the EUCalc Project. Available here
- Boris Thurm, Lucas Spierenburg and Marc Vielle (2018). Deliverable 6.1: Documentation on the GEMINI-E3 module and interface and on the way the library is generated. Public deliverable of the EUCalc Project. Available here
- Frank Vöhringer, Marc Vielle, Boris Thurm, Wolfgang Knoke, Dario Stocker, Anita Frehner, Sophie Maire and Philippe Thalmann (2017), Assessing the impacts of climate change for Switzerland. Final Report of CCImpact Project. Available here
- Daniel R. Klein, Ghazal Ebrahimi, Lucas Navilloz, Boris Thurm and Gunilla Öberg (2014). Water Management at UBC. Background report for the project: Would it make sense to develop an integrated resource management strategy for UBC, using a water lens? Available here

COURSES ATTENDED (Selection)

Microeconomics sequence, Prof. Klaus Schmidt, Prof. Piero Gottardi, Prof. John H. Moore, Prof. Jörgen Weibull, Swiss Program for Beginning Doctoral Students in Economics, Study Center Gerzensee
Macroeconomics sequence, Prof. Ricardo Reis, Prof. Sergio T. Rebelo, Prof. Fernando Alvarez, Prof. Jordi Galí, Swiss Program for Beginning Doctoral Students in Economics, Study Center Gerzensee
Environmental Economics, Prof. Philippe Thalmann, Dr. Vöhringer Frank, Dr. Vielle Marc, EPFL
Optimization and simulation, Prof. Michel Bierlaire, EPFL
Computable General Equilibrium in Climate and Energy Economics, Dr. Vöhringer Frank, Swiss Program in Environmental and Energy Economics, University of Bern
Financial Management of Energy Price Risk, Prof. Petter Bjerksund, Norwegian School of Economics, IAEE 2016 Summer School in Bergen
Introduction to Social and Economic Networks, Dr. Michael König, University of Zürich
Quantitative Models of International Trade, Prof. Samuel Kortum, Study Center Gerzensee
Long-Run, Global Macroeconomics, Prof. Per Krusell, Study Center Gerzensee

EXTRACURRICULAR ACTIVITIES AND INTERESTS

Nature lover, discovery of the world cultures and geography: Europe trip (Summer 2009), Peru and Bolivia (June 2012), China (Summer 2013), Canada and USA (April-September 2014), Central America (July 2015), South Africa (2017)

Sports: ice-hockey, played for 12 years, France champion U-16; mountain hiking; ski touring; climbing; squash; football

Sciences: Biology, Ecology, Hydrology, Thermodynamics, Quantum mechanics, Mathematics

SKILLS AND PERSONAL TRAITS

Language skills: French (native), English (business fluent), Spanish (spoken intermediate level), German (beginner)

Computer literacy: LaTeX, Matlab, KNIME, MS Office

Team spirit, eager to learn, proactive, flexible, open-minded, project management skills

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.