GERMINATING GOOD BEHAVIORS
A game prototype to test players’ incentive of choice

Master Degree Project in Informatics
One year Level 22.5 ECTS
Spring term 2019

Mikael Le Gal

Supervisor: Mikael Johannesson
Examiner: Jana Rambusch
Abstract

In this paper, we propose and develop a model for a prototype research tool based on a social dilemma game which use is widely spread in social psychology and experimental economics, the Public Goods Game. This tool generates from the necessity to expand the accessibility and versatility of this popular game as well as reach for newer audiences that might be otherwise deterred by the traditional Public Goods Game. Also, using this tool, we perform an experiment to try to find a possible preference towards either punishment or reward in the current population and find a possible correlation between said preference and three different cultural spheres (geographical, generational and gender).

**Keywords:** Cooperation, Public Goods Game, Human behavior, Serious Games, Game prototyping
# Table of Contents

1 Introduction 1

2 Background 2
   2.1 Social Dilemmas 2
   2.2 Cooperation 3
      2.2.1 Different basic behaviors 3
      2.2.2 Punishment, reward & reputation 4
   2.3 Cultural influences 7
   2.4 Humans vs AI 8

3 Problem 9
   3.1 Aim 10
   3.2 Methodology 10
      3.2.1 Procedure 10
      3.2.2 Participants 11
      3.2.3 Ethical considerations 11

4 Game iterations 13
   4.1 Game early prototype 13
   4.2 The final prototype version 15

5 Experiment 19
   5.1 Test pilot 19
   5.2 The final experiment method 20

6 Results and analysis 21
   6.1 Compilation of the data 21
   6.2 Survey data 21
   6.3 Classification of the test subjects 22
      6.3.1 Setting the age gaps 22
      6.3.2 Dividing cultures 23
      6.3.3 Gender classification 23
   6.4 Data analysis 24
      6.4.1 Contributions 24
      6.4.2 The carrot or the stick? 26
      6.4.3 Cultural differences 27

7 Conclusions 31
   7.1 Summary 31
   7.2 Discussion 31
   7.3 Future Work 32
   7.4 The game 33
   7.5 Societal-ethical considerations 34

References 35
Appendix A I
Appendix B II
Appendix C III
Appendix D IV
1 Introduction

Cooperation can be found with ease throughout nature, even between beings of different species. However, it becomes especially tricky when talking about human behavior. As rational beings we understand that acting selfishly has usually a better outcome (Sigmund, Hauert & Nowak, 2001), however some people still contribute selflessly when put into extreme situations similar to social dilemmas. Some even go one step further and assume a personal cost in order to try to correct selfish behavior (Fehr & Gächter 2002). Most studies on the subject have proven that punishment is much more effective than rewarding (Sigmund, Hauert and Nowak, 2001). However, when deciding to act on this, two possibilities arise: Will you punish the selfish in order to correct their behavior? Or will you reward the ones that cooperate so they continue to do so in the future?

In the following chapter, we will overview previous research relevant to this paper. We will start by defining the meaning of social dilemmas and set some examples, focusing on the Public Goods Game. Afterwards, we will discuss what cooperation is and its uses, as well as try to identify and categorize different human behaviors that can be seen when playing a social dilemma game. To finish the section we will see how some cultural differences might affect people’s behaviors when looking at them through the lens of cooperation.

Later on we will propose, design, develop and test a serious game based on the aforementioned Public Goods Game model, in order to use it for the experiment performed in this paper. Furthermore, we will set the aim of this paper trying to identify if there is a preferred incentive method, negative or positive, in the population and whether there is a correlation between that favored incentive and different cultural spheres such as the country of procedence, age or sex.

Afterwards we will overview the different iterations the game prototype has gone through, explaining its different changes and the reason behind them. We will also review the initial test pilot for the experiment part and come up with a polished experiment method.

Finally, we will analyze the gathered data in-depth and define the different cultural spheres we are about to use to classify the different test subjects and analyze the data through three different scopes. Then we will proceed to the conclusions as we summarize what we have learnt and try to make sense out of all the data gathered from a neutral perspective.
2 Background

2.1 Social Dilemmas

Social dilemmas are situations in which each member of a group has a clear and unambiguous incentive to make a choice that, when made by all members, provides poorer outcomes for all than they would have received if none had made the choice. Thus, by doing what seems individually reasonable and rational, people end up doing less well than they would have done if they had acted unreasonably or irrationally.

Dawes & Messick (2000, p.111)

Social dilemmas are interactions in groups of individuals where people who contribute to a public or common good would fare better in groups composed only of cooperators, as opposed to defectors, who do much better in mixed groups (Dawes 1980). Over the years, there has been a number of social dilemma games that have been used to experiment on the evolution of cooperation, such as the Prisoner’s Dilemma, the Snowdrift Dilemma and the Public Goods Game.

In Public Goods Games, a group of players receives a predefined initial income and afterwards they are able to invest any part of it in a common stock. After every player has made their decision, the common pot will be multiplied by a defined amount (it usually ranges from 2 to the amount of players) and then equally distributed among all of them. The payoff in classic PGG matches can be calculated using the formula below (Brañas-Garza and Espinosa, 2011):

\[ \pi_i = W - c_i + m \times \sum_{i=1}^{n} c_i \]

Where \( W \) meaning the initial funding or endowment, \( C \) the denoted player’s investment, \( M \) the factor that would multiply the common stock and \( n \) the number of participants. What makes this game so interesting is the fact that, without a proper way of controlling the behavior of the players, the temptation of not contributing anything and reaping benefit at the expense of the rest of the players becomes the norm, with most of the players becoming free riders (people that do not contribute) by the end of the game.
2.2 Cooperation

Evolutionary game theory—like the classical theory—predicts the selfish “rational” outcome. But if an arbitrarily small reputation effect is included in the analysis, a bifurcation of the dynamics allows for an outcome that is more “social” and closer to what is actually observed in experiments.

Sigmund, Hauert & Nowak (2001 p.10757)

For the purpose of this essay, it is important to define a precise way to identify and categorize cooperative conduct. As Van Lange and de Dreu (2001) explained in their paper, cooperation can be defined as a behavior that maximizes the outcomes of a collective. This collective, however, does not need to be composed entirely by beings of the same condition or even species (Dugatkin, 2002). To put it simply, any action that implies working on a specific matter knowingly or unintentionally benefiting other agents implied in it, independently of their differences or similarities, could be considered a cooperative behavior. As for the reason why these individuals feel the urge to join as a collective and cooperate, by-product mutualism theory merely attributes this drive to their will to increase their own individual fitnesses through cooperation (Brown, 1983). This desire to increase their odds through cooperation will rise when facing a common enemy, whether it is a physical threat or obstacle like a predator or a prey, or something incorporeal, e.g. harsh weather conditions (Mesterton-Gibbons and Dugatkin, 1992). There are other cases that could be classified as cooperative behavior, even if it’s as an accidental or unintended outcome. As explained by the private good model of by-product mutualism, any participant that generates a private good to himself but benefits other actors as a result of his/her actions, would fall into this model (Mesterton-Gibbons and Dugatkin, 1992).

Cooperation brings a series of benefits to the table. To begin with, it can increase someone’s self-confidence, as it has the potential of making you perceive yourself as helpful, and that may motivate you to continue with the cooperative behavior in future occasions, especially with a group with which you have collaborated before (Nelson and Norton, 2005). Hence, we can deduce that playing in a team against another team of players or AIs would generate a series of collaborative memories and cognitive links that may incentivize later collaboration with the same players. We can also assume the same way that direct competition should have the opposite effect as it may decrease future tentative of cooperation (Ewoldsen et al., 2012).

2.2.1 Different basic behaviors

If we take cooperation from a binary perspective, we can distinguish two kinds of actors, defectors and cooperators. Cooperators are players that partake in collaborative behavior while defectors are those who do not. However, that does not mean defectors cannot reap benefit from the cooperative players, even if they do not actively participate in a helpful way. De Jaegher (2018) presented the interesting example for cooperative hunting, where
defectors might still get a part of the prey if it is carried back to their location. If the target were consumed in the same spot, however, defectors would not get any benefit from the cooperative activity. Another curious example happens in collective defense, as a sentinel would not benefit from any other sentinels guarding the same area and only defectors would obtain by-product benefits (De Jaegher, 2018).

We should not forget the model of conditional cooperator, or conformist, either. A conditional cooperator modifies their own contribution to the collective depending on their own past experiences. This model is used to explain some behaviors during public goods games experiments, but could easily be extrapolated to other areas. A conditional cooperator might alter their attitude towards cooperation if they notice collaborative efforts are being rewarded, for example. It is also important to note, as Yang et al. (2018) did in their research, that some players might feel more inclined to raise their contributions faster than lowering them, as if they felt reluctant on taking advantage of the group.

So if defectors were sometimes still able to reap some benefit out of other people’s cooperation, what would make them change their conduct? It has been proven that in the public goods game, giving the option of punishment to community members encourages cooperation and slows down the end-game decline that is distinctive in ordinary public goods games. The opposite way of confronting this problem would be to let players reward helpful participants and, as stated by Yang et al. (2018) “endogenous rewards lead to a significant increase in cooperation compared with the control treatment and the trend of cooperation increases with rewards” (p. 9968). This leads to conclude that both systems promote cooperation, even if the effectiveness of such measures differ from system to system.

2.2.2 Punishment, reward & reputation
Being punished in a group of people can mean several things. Outside of mini-games or theoretical experiments, it can mean the rejection of the group, the loss of social status or even being physically threatened. In more controlled environments like the judiciary system or social experiments, it usually means a monetary compensation. In the latter case, that compensation often carries a minor consequence for the punisher who has to pay a cost to reduce the earnings of the sanctioned player (Barclay, 2006).

Following game theory, if the punisher has to pay a price as well, it usually means the behavior that leads to optimal individual benefit is not to punish at all and wait for other people in the group to do so. Nevertheless, there is still a considerable percentage of people that opt to lose some of their earnings in order to reprimand a selfish group member (Fehr & Gächter 2002). The main reason is that groups with people that punish bad behavior have higher cooperation levels than the ones that lack this kind of agent, resulting in a higher likeliness to stick together. In other words, the advantages of using
resources to correct other people’s conduct outweigh the negative effect that the punishers face, resulting in a general rise of the level of altruism. This kind of punishment, the one that is given regardless of the initial personal cost for the sake of the entire group, is called “altruistic punishment” and, as long as it benefits the team as a whole, it will tend to increase in the population (Boyd, Gintis, Bowles & Richerson, 2003).

Even if this kind of punishment is defined as “altruistic”, it could also carry several benefits for the punisher. A series of studies based on the Public Goods Game “demonstrated that people rated altruistic punishers as more trustworthy, group focused, and worthy of respect than non-punishers” (Barclay, 2006, p.341). Barclay (2006) also observed that, as long as the game was composed of enough rounds, the punisher’s reputation was increased among the rest of the components of the group due to their contribution correcting free riders’ behavior. Therefore, if the act of punishing a free rider is considered fair and the punisher perceived as a trustworthy individual, they might receive boons from other selfless agents. These benefits might come in the shape of more beneficial cooperative partnerships than those who do not punish. Consequently, being known for penalizing free riders could be advantageous as long as the defectors stop acting as such because they fear a possible sanction (Brandt, Hauert & Sigmund, 2003).

However, a system based in retaliation can also be taken advantage of or be used in a way that was not intended. Opposed to altruistic punishment we can find antisocial punishment directed by defectors targeting cooperators. In any population based on society, punishment is much less effective encouraging cooperation when these kind of antisocial strategies are involved (Powers et al, 2012). Hauser, Nowak & Rand (2018) argue that, if punishment can be targeted towards any subject, defector or cooperator, and given a group with a wide enough variety of components, people with all types of strategies would be present, including those that target cooperators when punishing. In this case, cooperators would be targeted to the same extent as defectors, thusly neutralizing the effect that altruistic punishment might have in said group and no longer stimulating cooperation. (Hauser, Nowak & Rand, 2018).

Basing cooperation on a reputation system can also be counterproductive depending on the culture. Cressman et al. (2013) found out during their research in China that in their experiments cooperation did not increase while using a punishing system, and a considerable amount of players chose to punish other members before knowing if they were prone to cooperating with them or not. Being these results vastly different from those obtained in western experiments, they conjectured that this difference was due to the different attitudes regarding reputation, being much less important in China to build a good reputation when a number of people interact repeatedly than in western societies.

As we have mentioned previously, positive incentives are meant to increase the contributors’ likeliness to cooperate in future exchanges by giving them some kind of reward. These kind of incentives are used in many levels outside of economic and sociological experiments and take many shapes and forms, like giving a present to a child
that has put a lot of effort in their studies or rewarding a team effort. On the other hand, these incentives are costly from an individual level for those who partake actively in this system, therefore being much more advantageous for those who do not (dos Santos, 2015). However, similarly to what we have already discussed with punishment and reputation, those investing their own welfare in positive incentives see their reputations increased towards those who receive said incentive. Not only that, but cooperation rate might also be increased based on the willingness to receive said reward, opposed to the fear of retaliation that we see in punishment systems (Gächter, Renner & Sefton, 2008).

Nonetheless, in the same way punishment can be taken advantage of and used as “antisocial punishment” against those who are contributing, there is an analogue version in reward mechanisms. Defectors can use the reward system to share the funds among them, thusly affecting cooperation even when posterior sanctions can be taken against these actors as this can lead to a society dominated by defectors or other agents that benefit from antisocial rewarding (dos Santos, 2015). Dos Santos (2015) predicted that getting rid of this problem depended on stopping free riders from being able to use the reward system at all and prosocial rewarders outcompeting them.

Reputation has a great effect in human behavior and that is reflected on social dilemma games such as the Ultimatum and the Public Goods Game. Reputation usually contributes to a fair exchange and incentivizes cooperation in any population that evolves through learning or imitation. When players such as altruistic punishers and prosocial rewarders contribute to the public good, the inclusion of reputation helps encourage the evolution of economically productive behavior (Sigmund, Hauert and Nowak, 2001). Social dilemma games such as the ones mentioned above, however, are often executed under the anonymity of their players, meaning they do not know each other before the experiment and they will probably not interact again after the exercise. This would arguably nullify the effects of any prior reputation a player might have. Still, we could also argue that, if a game is long enough for its players to interact repeatedly and start to comprehend and predict each other’s behavioral patterns, a micro-reputation could be built during that same match with its effects being the most apparent towards the end of the game.
2.3 Cultural influences

Up until now, we have reviewed the different possible general behavioral profiles we can find in any cooperative interaction, especially considering situations that are similar or directly related to Public Goods Games. However, it is not far-fetched to consider there might be external factors that may influence a person’s behavior or strategy in these kind of circumstances.

When it comes to gender difference, Vugt, Cremer and Janssen (2007) found out that men’s cooperation fluctuates greatly when there is conflict present within the group, while women contribute more in general but their behavior is largely unaltered whether there is conflict present or not, hinting at the possibility that men’s social behavior is more intergroup driven than women’s. Curiously enough, when observed by an in-group audience the behavioral gap between sexes is magnified. Under these circumstances, male individuals choose to be less cooperative while females do the opposite. A possible explanation for this difference is that men prefer their peers to see them as “tough” or “competitive” while women prefer to appear more cooperative (Charness and Rustichini, 2011). When using different tactics of influence such as joking, manipulating or complimenting, among others, there is not an ample difference between both genders. However, men tend to use certain tactics more often, such as manipulating the situation, joking and using the threat of punishment, while women prefer to make use of charm (DuBrin, 1991).

When considering age difference, some studies have found that age influences cooperative behavior. For example, young males that are more attractive cooperate less than their less attractive peers but this behavior becomes rarer when studying older males or females of any age (Shinada and Yamagishi, 2014). When testing the employees of two large firms, Charness and Villeval (2009) concluded that senior employees proved to be more cooperative than juniors were, even when there was a strong incentive to free ride on the contribution of the rest of the group. Another interesting detail from this experiment is that, when knowing they were participating with younger workers, seniors cooperated more than when participating in all-senior groups.

When making a comparison between different cultures, the results are not as clear. After performing the same experiment with subjects from Japan, the Netherlands, Spain and the USA, Brandts, Saijo and Schram (2004) concluded that there were only minor differences found in the behavior of the participants when comparing the four countries. In addition, they determined that in all four countries spite was not as strong a motivation as cooperation. In a different experiment performed in Austria, Japan and the USA, the results showed that unconditional cooperation was very similar between the three countries. However, the USA showed a much higher rate of conditional cooperation than the other two countries (Kocher et al., 2008). Finally, in a study made about 32 countries across the world, Harring (2016) attested that countries with low quality of government distrusted their fellow citizens, hence preferring punishment as an incentive, with the opposite case favoring a more rewarding system.
2.4 Humans vs AI

We have analyzed the possible human interactions and relationship evolution in a cooperative situation such as a social dilemma. However, would the different agents involved in this kind of environment behave the same way when faced with an artificial intelligence? Matters such as the actors’ reputation, cooperative behavior or trustworthiness would be of little importance for an AI, and the AI would not necessarily carry a reputation at all unless the human interacting with it knew what it was programmed to do beforehand. Studies such as the one conducted by Ravaja et al. (2005) indicate that playing against an avatar handled by a computer induces less spatial presence and engagement than playing against another human being. What is more, the players’ arousal ratings and physiological arousal are also higher when interacting with another person (Ravaja et al., 2005 and Lim & Reeves, 2010). Moreover, on a social level people tend to be less open, less agreeable, less extroverted, and less conscientious and self-disclosing, as well as showing a higher level of neuroticism when interacting with AI (Mou & Xu, 2017).

In order for a human being to perceive an avatar as being controlled by another player, there is a set of requirements that should be met. When studying the case to develop their AI model “Chameleon”, Tence et al. (2013) concocted a list of requirements for a computer-controlled avatar to be believable (table 1).

<table>
<thead>
<tr>
<th>Believability requirement</th>
<th>Summary of the requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>[B1: Reaction]</td>
<td>React to players and changes in the environment</td>
</tr>
<tr>
<td>[B2: Reaction time]</td>
<td>Simulate human-like reaction time</td>
</tr>
<tr>
<td>[B3: Variability]</td>
<td>Have some variability in the actions</td>
</tr>
<tr>
<td>[B4: Unpredictability]</td>
<td>Surprise the players with unpredictable behavior</td>
</tr>
<tr>
<td>[B5: Understandable]</td>
<td>Have an understandable behavior</td>
</tr>
<tr>
<td>[B6: Perception]</td>
<td>Have human-like perception</td>
</tr>
<tr>
<td>[B7: Planning]</td>
<td>Plan actions in the future to avoid mistakes</td>
</tr>
<tr>
<td>[B8: Memory]</td>
<td>Memorize information</td>
</tr>
<tr>
<td>[B9: Evolution]</td>
<td>Evolve to avoid repeating mistakes</td>
</tr>
<tr>
<td>[B10: Fast evolution]</td>
<td>Evolve fast enough for the players to see it</td>
</tr>
</tbody>
</table>

Table 1 List of believability requirements (Tence et al., 2013, p. 478)
3 Problem

Traditionally, Public Goods Games have always been performed in person, having to recur to mixing groups in order for the participants to be anonymous to one another. This also represents a problem when trying to make a generalized study of the behavior of the population since performing this tests in specific areas of said population keeps the subject base concentrated in one or a few chosen areas, which could be argued does not represent a whole population accurately. It is also sometimes inconvenient for the researcher to move to several locations in order to spread the subject diversity, and it is sometimes not possible due to time or economic constraints. Additionally, Public Goods Games experiments subjects have often a background in economics, which makes sense considering the game is a go-to in the field of experimental economics. However, it is a much skewed representation of the population and, in order to reach to a wider audience, must be made more accessible and its rules easier to understand for an inexperienced player base.

As mentioned in the background, there have been countless experiments using Public Goods Games and its myriad of different modalities. From traditional PGG without a method to try to prevent or stop free riders (Fischbacher, Gachter and Fehr, 2001 and Milinski et al., 2006), to more advanced modalities with punishment (Fehr and Gächter, 2000 and Nikiforakis and Normann, 2008) and reward (Szolnoki and Perc, 2010) only methods. In more recent years, there has been an increase in the amount of researches that combine both punishment and reward, but it is usually made optional (Rand et al., 2009) and the focus is put in the efficacy of these methods, instead on the predisposition of the population to choosing one over the other. Furthermore, as we have seen previously in the background section, there is a tendency for the players to become second-order free riders, opting for not acting at all on first-order free riding when given the chance to do so. As the author of this paper I want to raise the following questions: “When not given the chance to become a second-order free rider, what option to incentivize cooperation becomes natural to people?”, “Do any factors such as sex, age or culture influence this decision?”. As we have seen in chapter 2.3, differences such as sex, age and culture of origin seem to influence people’s cooperative behavior, so this paper aims to do something related and check if these differences also influence their incentive of choice.

To do so we would need a tool or game able to simulate one of these experiments while giving us enough freedom to set the ideal parameters to answer our questions. During the previous research for this paper, we have not found any such tool or game and, while some of them have a similar idea to what this paper proposes (Keil et al., 2017), they are lacking in other aspects that are needed for this thesis to come to fruition.
3.1 Aim

The main aim of this paper is to investigate if the current population has a preference regarding positive or negative incentives, and if there is any correlation between cultural factors and a possible prevalence towards choosing punishment or reward to correct other people’s behavior.

As stated previously, there are already studies that show how a difference in parameters such as age, country of origin and gender can mean a variance in cooperative behavior. It would not be too far-fetched to assume that those same factors might have an influence in people’s incentive of choice. Following this train of thought, the following hypothesis can be stated.

Hypothesis 1: The subjects tested will tend to use negative incentives, as they are seen as “more effective”.

Hypothesis 2: Cultural factors such as gender, country of origin and age do affect the way an individual prefers to correct other individuals’ behavior.

In order to solve these questions, a tool for research based in the Public Goods Game will have to be developed and tested.

3.2 Methodology

Although more data was gathered, the experiment focused only on one dependent variable: the players’ incentive of choice (the number of times an incentive was chosen over the other, from 0 to 10 times, with the possibility of skipping due to timeout). As for the independent variables involved in the results, those were the age, sex and country of origin of the participants. We have to consider as well other independent variables involved in the experiment that may have affected (or not) the results. Such variables would be the choices displayed every round, the allotted amount of time on each phase, the behavior set for the AI, etc. Such variables were constant through all sessions so all the subjects would play under the exact same conditions and can be found in appendixes A, B and C, which, combined with the accessibility of the game used, facilitates the replicability of the experiment.

3.2.1 Procedure

Since the game was played online, it was not possible to limit where the subjects could perform the test. Ideally, it would be in a location where the player felt comfortable and relaxed so they could focus on the task that was at hand and answer to the questionnaire rationally and in a detailed manner. Before starting the actual match, each participant went through the game’s tutorial screen where all the games’ mechanics were explained in a simple yet comprehensive way so the player would be able to take decisions based on
their own volition and not out of confusion. The test was repeated during several sessions, opening the server daily for a limited amount of time. While the server was open, the subjects were able to access the game and play against other players (that, unbeknownst to them, were controlled by the program). Previously to starting the match, the subjects were asked to give some background information such as their age, sex and country of precedence. During the match, the game registered a set of quantitative parameters relevant to the experiment, such as the players’ contributions and defections, the time they took to act and who they chose to punish or reward for every round (see Appendix E for more details). At the end of every round, the game would send such data to the server, which registered it in a document. At the end of the match, the subjects were prompted with a questionnaire and a similar process would take place to register their answers. The information from the questionnaire was used mainly to receive some feedback regarding the game’s aesthetics and mechanics and to test if the subject answering the questionnaire believed they were playing against human players or not. If there were clear enough indications that the subject did not believe the rest of the players were human, that subject’s data would be eliminated.

3.2.2 Participants

Due to time constraints and a limited amount of test volunteers, the participants for the experiment section were chosen through “snowball sampling” (also known as “chain sampling”), meaning the first stage of participants was selected by the researcher and then those participants selected the subjects of the next stage among their acquaintances, and so on (Goodman, 1961). Following standard PGG procedure and for the effects of this study, the participants were kept anonymous. There were no restrictions for participation, except having access to a computer and being able to operate it. The original intent of this paper was to represent a segment of the population as wide as possible (thus the lack of restrictions to participate). However, due to time constraints there was a limited amount of participants. Twenty-four people participated voluntarily, their age ranging from 23 to 66 years old, most of them European citizens (France, Germany, Spain and Sweden). Fifteen of the participants were males, while the rest were females (no participant considered themselves non-binary).

3.2.3 Ethical considerations

After being informed about the general implications of the experiment and the general characteristics of the game and how to proceed, the players were asked to participate voluntarily, in other words: informed consent was used. To ensure their anonymity was kept, they were not asked to input their actual names with the rest of their personal data (but a nickname). This ensured confidentiality for the first stage of participants (since the researcher directly contacted them) and complete anonymity for the following stages. The data handled was secured and kept from being displayed publicly. The game presented no potential harm for the participants and the researcher had no ulterior motive to
perform this experiment besides the ones stated in this paper. (Lune, H. and Berg, B.L., 2016, pp. 43-64).
4 Game iterations

As mentioned in the aim section (chapter 3.1), a game prototype had to be designed, developed and subsequently tested in this paper. The game would have to adhere to the basic rules of Public Goods Games, although altering some of them to make the product more entertaining, accessible and easier to play. To access a wider audience, the game would have to be playable online with the server registering the players’ choices in order to gather data for the research part of the thesis. The game would be playable on computers (Windows OS) and only a keyboard and mouse would be needed to fill the data and play the game. The game would have to include a comprehensive tutorial considering some of the subjects could lack, potentially, any experience playing Public Goods Games or video games in general prior to the experiment.

Due to the limited amount of time, and to make the matches comparable between each other, the game lost its multiplayer aspect with 3 out of 4 players being controlled by the program itself in a deterministic manner. This would allow the researcher to compare directly the subjects’ behavior since all the parameters but the player’s choices would be identical in all matches. Of course, as was mentioned in the background section, a human player would act differently when facing an opponent controlled by an artificial intelligence so the game had to simulate human behavior in order to fool the player. Something similar was attempted in the Pizza Game, although the objective public were children and adolescents from 9 to 16 years old (Keil et al., 2017). In order to make players believe they were playing against human opponents the game would have to adhere to the basic requirements for AI believability mentioned earlier (see table 1). Considering the game was turn-based, requirements such as reaction, reaction time and perception were easy to mimic. However, we would have to make a compromise regarding other aspects such as planning, memory and specially evolution. For the same reasons stated above regarding the involvement of other human players, if we designed a non-deterministic AI that adapted and evolved to the player’s actions we would no longer have identical matches to compare and this “noise” would have made it harder to reach a conclusion.

4.1 Game early prototype

The game is a modified version of the traditional Public Goods Game (PGG from now on) with added reward and punishment systems. As in the classic PGG, the basic rules are as follow: players participate anonymously in the game. Each one of them starts the game with a set amount of points that they can invest secretly in a common pot. Afterwards, that pot is multiplied by a predetermined factor (usually between 2 and the total amount of players) and split evenly between all the players, even the ones that have not invested any points in the pot.

The difference with the standard PGG, besides the punishment and reward system that will be explained later, is that in this modified version players have limited options for investing their points. The game is played by 4 players (1 human and 3 controlled by the
program), lasts for 10 rounds and every round is divided in two parts. Each part has a timer to control how long players take to make decisions.

In the first phase, the birds have to collect energy from flowers that randomly spawn in a cherry blossom tree. In order to do so, they spend some of their existing energy to reach those flowers. Depending on what height the flower is located, the player will have to invest more or less energy to reach it, but higher flowers also contain higher values of energy (see figure 1). Once selected, birds will consume half of the flower’s corresponding energy, and the other half will be added to a common pot.

![Figure 1](image1.png)  Early first phase

On the second part of each round, players are given the option to punish or reward another player. Both options cost energy and have different effects. Punishing is very costly, but also leaves the player chosen without the option to partake in the common pot this round. Rewarding a player will make their energy cost when moving in the first part of each round more efficient, effectively making their investment more profitable. There is no way to opt out of choosing one of these two options, although if the timer runs out without choosing an option the player will be sanctioned and some of their points will be taken away anyways (see figure 2). After every player makes their decision, their choice will be communicated to the rest (see figure 3) and then the pot will be split evenly among the players that have not been punished.

![Figure 2](image2.png)  Early incentive choice
The similarities with the standard PGG are still there as players still invest points (energy) when their avatar flies to reach the flower they selected and those points are multiplied by a factor (the flowers are always worth more than the energy invested to reach them). After they either punish or reward another player, the points are evenly divided among all the players (except the ones that have been punished, if any). The basics of any PGG are still present in this version, but there is some added randomization and strategic choices to be made, which adds some depth to the game, thus making it presumably more enjoyable and rewarding for the players.

After the game, a questionnaire is prompted on the screen to collect some qualitative extra data and feedback from the players. The questions range from the player’s opinion about the rest of the actors in the game to more game specific questions like their reason to choose an option over another, their strategic decisions and why did they decided to support or punish a specific player. As the questionnaire is automated, this allows for more tailored questions, varying from player to player depending on how they played throughout the game.

The game has imbedded an auto-log feature that registers every player’s choice during the game, as well as summarizes the information collected for easier post-analysis and also registers the post-game questionnaire and player data. This could enable the game to be uploaded and played online without the need for supervision during the match, although this was not implemented yet in this version of the prototype.

4.2 The final prototype version
The version of the prototype used in the experiment features several changes and additions to the original formula. Following the feedback received during and after the original experiment pilot, some of the game’s rules and core mechanics were simplified or completely changed in order to make it more accessible to new players or people that are not used to playing games that are more complicated. The following changes or additions were made:
- The game was made playable online through a Windows application. Originally, the idea was to make the game playable on any internet browser but, after having several problems trying to host the game, the researcher resolved to a less convenient but more reliable solution.

- The core rules were modified as follows (refer to chapter 4.1 to compare): after the players choose the flower they want to contribute with (or not contribute at all) the total amount of points will be added to the common stock and split evenly among all players right after. After that, the option to punish or reward will be prompted on the screen. Punishing another player will take 25% of the points earned this round from the punisher, and 50% of the points earned this round from the objective player. Rewarding, in a similar way, will take 25% of the points earned this round from the player that is rewarding and add an extra 50% to the objective player. The reason these core mechanics were modified was to try to make both options seem more even or balanced, as well as to make the rules easier to understand and closer to the standard Public Goods Game.

- A server application was made in order for the researcher (and potential future users) to customize the game’s many variables easily and without having to modify the core game’s scripts every time. The server app features an array of different fields or variables to modify, from the amount and value of the flowers each round to the AI’s behavior and questionnaire prompts (see figures 4 and 5).

![Server main screen](image1)

**Figure 4** Server main screen

![Server point distribution settings](image2)

**Figure 5** Server point distribution settings
- A tutorial splash screen (figure 6) will teach the players the game’s different rules and mechanics right before starting the match to avoid any possible confusion during the test.

![Tutorial splash screen](image)

**Figure 6** Tutorial splash screen

- To make every test’s conditions identical, the flowers no longer will spawn in a random quantity and with random values (confined within a range). Instead, they will spawn in the exact amount and with the exact values specified in the server program.
- For the same reason, the AI will no longer try to adapt to the player’s decisions. Alternatively, it will proceed as indicated in the server program (becoming deterministic).
- The questions prompted in the final questionnaire will no longer be hard-coded in the game but rather input through the server program. Although this will make it impossible to pinpoint and ask about a player’s specific actions during the test, it will make it easier for future potential researchers to add or modify the prompts of the questionnaire on a whim.
- A lobby simulating a matchmaking in process and some delays simulating other players input were added in order to cover the believability requirements suggested by Tence et al. (2013).
- Some animations, particle effects and splash screens were added in order to make the game more visually pleasing, enjoyable and easier to understand (figure 7).
The auto-log no longer presents a “wall of text”. In this version of the prototype it splits the in-game data and the questionnaire into two different types of document. The former is stored as a table in an *Excel* document and the latter is stored as a text document. This makes it much easier to access to the different kinds of data and, in the case of the excel document, it also makes it easier to process the data through excel’s calculation features.

After implementing the aforementioned features and changes, the prototype was tested and its client part uploaded for posterior sharing with the possible test subjects.
5 Experiment

5.1 Test pilot

To test this experiment, a pilot was performed with the early unfinished version of the prototype (the one described in chapter 4.1), as the game was still on development. However, basic functions and game mechanics were already implemented. The test was performed in person on two different subjects. Because the tutorial screen and other features were still missing, the researcher gave a basic explanation of the game mechanics and the subjects were able to practice a little in a practice mode to make sure they understood them.

After that, a normal match was started with little to no further explanation about the game. Everything performed as expected including the automatic register of the match progression with the exception of the players controlled by the program, which presented an unexpected behavior during the game. That probably hindered the illusion of playing a game against other human players along with the lack of other parts of the program that were still in development by the time the pilot was tested, dedicated to simulate an online match.

After finishing the game and completing the questionnaire, the players were further questioned about the game’s performance and design, and the program-controlled players. Both of them did not find the situation believable, among other reasons because of “the context of the situation”. In future iterations of the game, more emphasis had to be made in the fake online matchmaking to make sure players accepted the possibility that they were playing against other humans. The game was generally perceived as “entertaining”, although in a couple of occasions the subjects were confused about the mechanics. The log, although extremely detailed and mostly helpful in its early stage, did become a hindrance because of the way the information was registered, as it presented a “text wall”.

Furthermore, although the results were inconclusive due to the low number of participants, their main reason to use a choice over the other was mostly strategic. Despite the fact that both of them selected the reward option much more often than the other, it was while they were ahead of the competition. As stated by one of the subjects, “If I were losing I would have chosen the punishment option over the reward”. Although these results would still be valid for analysis, both options (punishment and reward) would have to be re-balanced in a later version so players would choose them out of personal preference and not because one was clearly more advantageous than the other. Further development and fine-tuning had to be made on the core mechanics of the game before the final version was ready to be used for the experiment.
5.2 The final experiment method

Once the prototype was uploaded and ready to use, it was handed to the first stage of participants. They were given brief instructions about how to download and install the game, as well as a brief summary of the purpose of the game. Then they were asked to suggest the experiment to other acquaintances of theirs in order to gather more subjects, as the snowball sampling method proposes. The server was installed in a personal computer and its parameters for the experiment set. There were 8 sessions, with 24 subjects in total.
6 Results and analysis

6.1 Compilation of the data
The data gathered from each player’s match covered a wide range of information: player’s age, sex and country of origin, the investment made on each round as well as the contribution gathered, the incentive chosen on each round (punish, reward or timed out), the time invested choosing an option on phases 1 and 2 as well as the total amount of points at the end of each round. Also, the answers that were input in the post-game questionnaire. All this data was put together, summarized and put into a table for posterior analysis (see Appendix E). From the information accumulated, age, sex and country of origin was used to classify and categorize the different subjects so their results could be compared depending on their generation, cultural region or gender. As for the time taken to take a decision, it served as a filter to estimate if players were carefully considering their options or mindlessly clicking. Seeing that the options in the game are never too many or too difficult to take, the minimum average amount of time to pass this filter was set to 4 seconds for the first round and 3 seconds for the second phase.

6.2 Survey data
As stated previously, the survey data had two objectives: the first one was to gather feedback about the game itself so new iterations could improve on the previous concept and make the game more enjoyable. The players had a mixed opinion about the game. When asked “Did you like the game? What did/didn’t you like about it?” some of them perceived it as fun, or interesting, i.e. “I like this game because you have to guess/bet what other players’ the next moves are going to be. I think it would be a fun/useful game to explore one’s mental decision making ability for research purpose” or “Funny! Cool animations and sounds. I would love to have emotes/change the personalty of my bird” while others found the game to be boring or the rules too complicated, i.e. “I didn’t really understand it quick enough. Also, the acting to waiting ratio is a little bad (too much waiting, too little acting)”. With this information gathered, future iterations of the game could have improved mechanics to make it less confusing or improve the wait to act ratio.

The second purpose of the survey was to identify possible players that did not believe they were playing against other human beings and rule them out of the data. In order to do that, the questions had to avoid being too direct, since there was the possibility that some players would lie not to admit the program had fooled them. To achieve that, the questionnaire was designed to go back and forth between questions related to the match and questions to test the player’s opinion on the rest of the players (see appendix D to see the questions). No players showed a clear sign of disbelief against the fact that they were playing online against other players. Some of them even personalized or antagonized the birds controlled by the program, i.e. “Some were lazy pickers! I got the feeling that they played inconsistently, but thats hard to say given I don’t know the best practice. I think
they tended to reward rather than punish, until the last rounds, when their true, darker, colors appeared.” or “I think the yellow bird didn’t like me”.

6.3 Classification of the test subjects

6.3.1 Setting the age gaps

The age range of the test subjects starts from a minimum of 23 years old and tops at 66 years old, with the bulk of the participants concentrating around the late twenties and early thirties. There are a myriad of factors to consider when trying to divide a human’s lifespan in small meaningful age ranges, and those factors can vary from culture to culture. For example, while in some places emancipation usually takes place between 18 and 22 years old, due to economical crisis and cultural differences some young adults, especially around the southern European region, still live in their parents household at around 30 years old. However, in order to be able to define those different age ranges, we will have to make a compromise and try to reach for an understandable middle ground between the different cultures.

Erikson’s (1982) life-cycle theory divided any human’s lifespan in 8 different stages attributing a specific psychological issue to each one of them and splitting adulthood into 3: young adulthood (from 19 to 40 years old), middle adulthood (from 40 to 65 years old) and late adulthood (from 65 years old onwards). There’s some critique to Erikson’s work, however, as his theory seems to be centered on European-American males (Gilligan, 1982) and “for lacking sufficient complexity to represent the epigenetic mechanisms involved” (Sacco, 2013). We will use the updated version of Erikson’s eight stages of life suggested by Sacco (2013): the Fibonacci Life-Chart Method.

The Fibonacci Life-Chart Method (FLCM) provides a biopsychological basis for Erikson’s life-cycle theory and eight stages founded on the occurrence of Fibonacci numbers in biological cell division and self-organizing systems.

Sacco (2013, p.145)

This method also divides adulthood in three different stages: young adulthood (from 18 to 29 years old), middle adulthood (from 29 to 48 years old) and older adulthood (from 48 onwards). With this in consideration we can split our participants into the different three stages as seen in the next table (table 2):
Table 2  Participants per age division

<table>
<thead>
<tr>
<th>Age division</th>
<th>Number of participants</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young adult (18-29)</td>
<td>12</td>
<td>50%</td>
</tr>
<tr>
<td>Middle adult (30-48)</td>
<td>10</td>
<td>41.67%</td>
</tr>
<tr>
<td>Older adult (49+)</td>
<td>2</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

6.3.2 Dividing cultures

The vast majority of the test participants (22 out of 24) originate from within Europe, with the rest belonging to China and Argentina respectively. Considering that there is not enough representation for the two latter players, we are going to focus our efforts in dividing the European continent into logical and comprehensive regions. There are multiple approaches to this problem, though, as we can consider geographical, cultural or political-economic boundaries. For the sake of simplicity and considering there is an ample variety of cultures within Europe, thus making the cultural classification over complicated or even useless, we are going to apply the geographical method. The United Nations Statistics Division (2019) subdivides the European continent into 4 large areas: Eastern Europe, Northern Europe, Western Europe and Southern Europe.

Excluding the 2 participants that have their origins outside of Europe, we can find subjects in our data belonging to Northern Europe (Sweden), Western Europe (France and Germany) and Southern Europe (Spain). Once organized, the three groups would be distributed as follows (table 3):

Table 3  Number of participants per geographical region

<table>
<thead>
<tr>
<th>Region division</th>
<th>Number of participants</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northern Europe</td>
<td>7</td>
<td>29.17%</td>
</tr>
<tr>
<td>Western Europe</td>
<td>7</td>
<td>29.17%</td>
</tr>
<tr>
<td>Southern Europe</td>
<td>8</td>
<td>33.33%</td>
</tr>
<tr>
<td>Others</td>
<td>2</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

6.3.3 Gender classification

In the beginning of the play session, the test subjects are requested to input their gender from a dropdown menu with three possible options: female, male and non-binary. Out of the 24 final participants, none considered themselves “non-binary” thus we are left with just the other 2 groups. After dividing the players into both categories we are left with the following table (table 4):
<table>
<thead>
<tr>
<th>Gender division</th>
<th>Number of participants</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>9</td>
<td>37.50%</td>
</tr>
<tr>
<td>Male</td>
<td>15</td>
<td>62.50%</td>
</tr>
</tbody>
</table>

Table 4  Number of participants per gender

6.4 Data analysis

6.4.1 Contributions
Contributions refer to the amount of points each player committed to the common pot at the end of the first phase in every round. This data, paired with the investment linked to this contribution, helps categorize the different players’ behavior. It is NOT the aim of this paper to make any comment or statement about the possible link between contribution and the different cultural factors analysed in this category. However, it is still important to evaluate a possible relation between the player’s cooperative behavior and their incentive of choice. The following box plots represent the average contribution per turn of the players when separated by age, country or gender.

When compared by age, we can see that, although having similar extremes, the younger players studied in this experiment tended to invest less than those in the range of “middle adult”. As we can observe (see figure 8), the “young adult” box plot is situated relatively lower than the “middle adult” one while still keeping a similar size, meaning they defected considerably more times than the others. The difference between the mean of the young adult group (281.58±136.54 points, n=12) and the mean of the middle adult group (329.50±126.56 points, n=10) was, however, not statistically significant (twosample t-test, p = 0.41 > 0.05).

![Figure 8](image)

Figure 8  Average contribution per age range
When separated by region, we can notice wider differences in the graph; especially if we compare the western European region with the other two (figure 9). While the southern European subjects' box plot is located further down than the northern European one, its spread is more concentrated and both regions’ median and average are very similar; meaning both regions had a similar cooperative behavior, albeit the northern European one showed a wider gamut of choices. However, the western European players have clearly contributed much less in average than the other two regions, having its median below the 200 points range. The difference between the means of the Northern European (329.57±154.67 points, n=7), Western European (228.71±162.59 points, n=7) and Southern European (331.50±122.86 points, n=8) regions were found also not statistically significant (twosample t-tests, Northern-Western: p = 0.26 > 0.05; Northern-Southern: p = 0.98 > 0.05; Western-Southern: p = 0.19 > 0.05).

![Figure 9](image)

**Figure 9** Average contribution per geographical region

Finally, the difference between genders is also noticeable in the graphs (see figure 10). While both show a similar value for their third quartiles (meaning that values below that range represent 75% of the players in that group), the female group has higher, more concentrated values and it shows a median about 100 points above the male group; connoting that the women participating in this experiment were more prone to cooperating than defecting or at least more keen on contributing points for the rest of the players, as opposed to the men that played the same game. Then again, the difference between the mean of the male group (280.87±147.18 points, n=15) and the female group (337.67±132.37 points, n=9) were not significant (twosample t-test, p = 0.35 > 0.05).
6.4.2 The carrot or the stick?
As described in previous sections, every round after choosing their contributions to the common pot players engaged in deciding whether to punish or to reward another player without having an option to pass (letting the timer run out would result in the loss of points without carrying any additional benefit). This design choice was made to avoid the appearance of second-order free riders, who would make this case study significantly harder to measure.

In order to determine if a player showed a clear preference for one of the two incentives, the threshold was arbitrarily set to 70%. In other words, for this study to consider that a player has a prevalence, that player would have to choose the same option at least 7 out of 10 rounds. The reason this proportion was chosen was because it strayed far enough from the medium point of 50% while not being a statistical rarity. Once this was set and players with a mixed tendency were identified, we could proceed to compare the data with our hypothesis.

We can deduce the following cases:

Case 0: The three groups (punisher, rewarder and mixed behavior) are equal.

Case 1: Most players prefer to punish bad behavior over rewarding a good one.

Case 2: Most players feel compelled to reward good behavior rather than punishing a bad one.
Once the parameters are properly set, we can proceed to analyze the data gathered during the experiment. From the 24 players that participated, nine showed a prevalence towards punishment, while seven preferred to reward. The eight players left did not show a clear preference between negative incentives and positive ones, and their behavior was classified as “mixed” (see table 5 for proportions).

<table>
<thead>
<tr>
<th>Type of incentive preference</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punishment</td>
<td>37.50%</td>
</tr>
<tr>
<td>Reward</td>
<td>29.17%</td>
</tr>
<tr>
<td>Mixed</td>
<td>33.33%</td>
</tr>
</tbody>
</table>

Table 5  Global preferred incentive

As we can see in the table above (table 5), there is some difference in numbers between the amount of participants that preferred punishment and those who preferred to reward. However, this difference was not significant ($\chi^2_{0.05}(2, n = 24) = 0.25, p = 0.88 > 0.05$).

### 6.4.3 Cultural differences

To try to answer the second question “Do cultural differences, whether geographical, generational or sexual, influence the preferred incentive?” we proceed to analyze the data collected with more thoroughness, separating the test subjects in the aforementioned groups.

A chi-square test of independence will be run for each population group in order to determine if there is an association between the variables or, simply put, if the incentive of choice is actually correlated to age, country of origin or sex.

If we take a look at the graph below (figure 11), we can notice some progression in the players behavior as they grow older. Older players that participated in this experiment show more predisposition to using positive incentives than punishing other players. We do have to note, however, that the participants belonging to the “older adult” range represented only 8.33% of the total amount of subjects so their results, although interesting when looking at the entire scope of age ranges, do not accurately represent the entirety of their group.
Figure 11  Median of times an incentive was selected, by age range.
Box = 1st and 3rd quartiles; bars = min and max values; orange dot = average.

After classifying the subjects in the three aforementioned behavioral groups (punishers, rewarders and mixed behavior), a chi-square test was run in order to pinpoint a possible correlation between age range and said behavior. Although we can appreciate in the graph that the amount of times an incentive was chosen slowly becomes the opposite with time, this difference was found not to be significant ($\chi^2_{0.05}(2, n = 24) = 4.56, p = 0.34 > 0.05$). Thus, the null hypothesis cannot be rejected. That is, there is no relation between the variables.

Regarding geographical cultural differences, western and southern European players in this experiment show a similar behavior when choosing an incentive, both respectively boasting a median of 8 and 7 times punished of a maximum of 10, an a low amount of rewards, although the numbers were more spread out in the Southern European group. Northern Europe shows a more balanced behavior with rewards being chosen generally more than punishments (figure 12).
Figure 12  Median of times an incentive was selected, by geographical region.
Box = 1st and 3rd quartiles; bars = min and max values; orange dot = average.

After classifying the different regions and running the chi-square test of independence, the p-value obtained is 0.038 ($\chi^2_{0.05}(6, n = 22) = 13.33, p = 0.04 < 0.05$). As it falls below the significance level chosen (0.05), we could reject the null hypothesis and assume that there might be a correlation between country of origin and incentive of choice.

When analyzing the relation between gender and the incentives chosen, we can see that the male players that participated in this experiment showed a somewhat aggressive style of play, opting for lower contributions and punishing much more than their feminine counterpart. If we observe the graphs below (figure 13 and table 6) we can appreciate some prevalence of punishment over reward, next to mixed behaviors. Only two out of 15 male players chose to reward at least a 70% of the times during the match.
Females that participated in this experiment showed more extreme incentive predilections, with more than half choosing the same incentive at least 80% of the times. A majority of female players also chose reward over punishment or mixed options (figure 13 and table 7).

Following the same procedure as in the last two groups, if we classify the subjects based on their incentive of choice and run a chi-square test, the p-value obtained is still higher than the significance level ($\chi^2_{0.05}(2, n = 24) = 3.48, p = 0.18 > 0.05$). Thus, we must conclude that there is not enough evidence to suggest a possible association between gender and the incentive of choice.
7 Conclusions

7.1 Summary

Two questions were formulated in this paper: “When not given the chance to become a second-order free rider, what option to incentivize cooperation becomes natural to people?” and “Do any factors such as sex, age or culture influence this decision?” As we could see in chapter 6.4.2, there was not enough evidence to suggest that there is a preference towards one incentive over the other. Thus, we must reject our first hypothesis (Hypothesis 1: The subjects tested will tend to use negative incentives, as they are seen as “more effective”) in favor of the null hypothesis (There is not a preferred incentive of choice).

As for the second question, in chapter 6.4.3 we proved that factors such as age and gender do not affect this decision. However, a possible correlation between the incentive of choice and the geographical origin of the participants was found. Therefore, if we consider the second hypothesis (Hypothesis 2: Cultural factors such as gender, country of origin and age do affect the way an individual prefers to correct other individuals’ behavior), its null hypothesis (Cultural factors such as gender, country of origin and age do NOT affect the way an individual prefers to correct other individuals’ behavior) cannot be completely rejected as that could cause a type I error (false positive). Following the results, we have to reject that gender and age affect the way an individual prefers an incentive to another. The influence of the country of origin, however, showed a statistically significant result and cannot be rejected.

7.2 Discussion

The first thing we must discuss is the validity of this experiment. Regarding the internal validity, we cannot be completely certain that the independent variables are the sole cause for the behavioral patterns, as there could be other minor variables that were not considered affecting the results. For example, for the second part of the experiment we analyzed the same data under 3 different lenses: age, region and gender. While the researcher of this paper meant to keep things simple, it is important to note that there are a myriad of different factors besides the ones that were used that might be directly or indirectly responsible for the statistical differences. Regarding the positive results related to the country of origin, it has to be mentioned that these results might be due to a possible imbalance in the sample, as all of the Western European representatives were male (while the other regions were more balanced). That being said, no relevant results were found for gender differences. As for the external validity, this paper focused on the Public Goods Game, which falls under the category of a social dilemma game. This means that the players’ behavior could probably be similar in other social dilemma games. However, we cannot guarantee that the results observed here would translate directly to real life decisions, as some players might have preferred to punish others just because they conceived it as a valid strategy and not for personal reasons, for example.
For time constraint reasons, a split half method could not be performed in order to test the experiment’s internal reliability. As for the external reliability, the tool used records every decision taken with precision and was tested in the pilot prior to the final experiment. Furthermore, the game used in this experiment is publicly available (see section 7.5 for more details about it) and the exact parameters used in said tool can be found in appendixes A, B and C; so the experiment can be easily replicated and tested.

It is important to establish whether the findings gathered here in this paper can be used directly or not. The scope of the experiment was considerably big, since the aim was to analyze a population as general as possible. While the data herein might be of some use, the total amount of subjects is not enough to represent accurately the general population of Europe, even less the Global population. In addition, since the sample was relatively small, we might find some important deviations if we replicate this experiment, hence the importance to do so with a much bigger subject pool and amount of time.

Even so, although the results must be taken with a grain of salt, we could still take them as a hint or indication of a possible difference between cultural factors.

7.3 Future Work

Regarding the prototype developed in this project, and considering the feedback gathered through the questionnaire and after performing the experiment, some changes or upgrades would be in order before using the game in another experiment:
- The game needs to be optimized to avoid crashes in the future.
- The game could use some kind of interactive tutorial, for those users that have trouble understanding the rules.
- Compatibility with other platforms could be added, although it is not mandatory (some volunteers could not participate in the experiment due to owning a non compatible system).
- Integrating real multiplayer would enhance the game’s versatility. The game is already built around a multiplayer system, but it needs more development time to fully integrate the feature.

As for the research found in this paper, and considering the results are not entirely conclusive for the reasons previously stated, this study would greatly benefit from a replication with a much larger subject base and an updated prototype. Thanks to the already existing prototype, this should not prove to be too difficult (except for the gathering of test volunteers) since the game could be reused and the parameters used for this experiment can be found in the Appendix A, B and C.

Another possible project could spawn from this study if we consider the possible correlation between gender and the incentive of preference. Using the same prototype, or an upgraded version, and the same or completely different parameters for it, we could
sample subjects from the same geographical area and age gap and then focus only in the possible differences found between the various genders. If any definitive correlation is found after performing said experiment, we could try to delve deeper into the subject and try to establish the reason of said difference, whether physiological, educational, environmental or any other. Of course such a study would need to be composed of an interdisciplinary team, since such matters escape the field of expertise of the researcher of this paper.

Another future study could be analyzing the possible correlation between high contributors and a preference towards positive incentives (rewards). When comparing the results in chapter 6.4.3 with the ones related to contributions (chapter 6.4.1), we can see that there might be a correlation between cooperative behaviors and a certain prevalence for an incentive type as we can see that, in the majority of cases, groups that contributed the most also showed certain preference for choosing reward over punishment.

7.4 The game

For this paper we proposed, designed and developed a prototype of a game based on the Public Goods Game with punish and reward mechanics. The game is meant as a tool for future research, as its parameters are fully customizable and the game is playable online which should make the experiment more accessible to a general public. During the experiment performed in this paper, the game was proven to be successful for what it was designed for. Under optimal to good conditions the game worked flawlessly and delivered the data to the server, where it could be analyzed afterwards. Under suboptimal conditions, however, several crashes were reported presumably due to an unstable internet connection or a lack of optimization in most of the cases. Crashes and other issues are to be expected, since this is a prototype and by no means a finished product.

It is not the intention of the author of this paper to gain any profit from the distribution of the game but to share it publicly so any researchers interested might be able to use it for their own purposes.

Although the game is still an early prototype, the author of this paper does not compromise to continue the development any further than what he is willing to do. However, the game is open source and its script can be modified and/or improved by anyone that might be interested.

The game can be requested directly to the author of this paper through e-mail:

legal.mikael@gmail.com
7.5 Societal-ethical considerations

It is, by no means, the objective of this paper to cause segregation or fuel any kind of prejudice towards any group of the population studied. While the geographical differences were found relevant and a group showed to be more lenient and favored cooperation and reward more than others, this experiment is based on a social dilemma game and cannot and should not be directly translated to other real life situations, as it was already stated in chapter 7.2. Instead, the content of the experiment should be used to try to discern, in future research, the origins of such a difference and try to find the means to change, if deemed necessary, an overly competitive or strict behavior on a certain group thus benefitting instead of harming a certain population.
References


## Appendix A - Experiment main and incentive settings

<table>
<thead>
<tr>
<th>Settings</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of human players</td>
<td>1</td>
</tr>
<tr>
<td>Number of AI players</td>
<td>3</td>
</tr>
<tr>
<td>Number of rounds</td>
<td>10</td>
</tr>
<tr>
<td>Timer (1st round/2nd round)</td>
<td>30/20</td>
</tr>
<tr>
<td>Reward option</td>
<td>On</td>
</tr>
<tr>
<td>Punishment option</td>
<td>On</td>
</tr>
<tr>
<td>Skip option</td>
<td>Off</td>
</tr>
<tr>
<td>Incentive multiplier</td>
<td>2</td>
</tr>
</tbody>
</table>
## Appendix B - Experiment flower (points) settings

<table>
<thead>
<tr>
<th>Branch 1</th>
<th>Branch 2</th>
<th>Branch 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Round</th>
<th>F1 (value/branch)</th>
<th>F2 (value/branch)</th>
<th>F3 (value/branch)</th>
<th>F4 (value/branch)</th>
<th>F5 (value/branch)</th>
<th>F6 (value/branch)</th>
<th>F7 (value/branch)</th>
<th>F8 (value/branch)</th>
<th>F9 (value/branch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>100/1</td>
<td>160/1</td>
<td>200/2</td>
<td>280/2</td>
<td>400/3</td>
<td>600/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 2</td>
<td>100/1</td>
<td>120/1</td>
<td>200/2</td>
<td>400/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 3</td>
<td>160/1</td>
<td>200/2</td>
<td>240/2</td>
<td>400/3</td>
<td>480/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 4</td>
<td>80/1</td>
<td>100/1</td>
<td>120/1</td>
<td>200/2</td>
<td>260/2</td>
<td>400/3</td>
<td>520/3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 5</td>
<td>2000/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 6</td>
<td>100/1</td>
<td>160/1</td>
<td>220/2</td>
<td>280/2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 7</td>
<td>100/1</td>
<td>200/2</td>
<td>360/3</td>
<td>400/3</td>
<td>440/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 8</td>
<td>100/1</td>
<td>120/1</td>
<td>140/1</td>
<td>200/2</td>
<td>240/2</td>
<td>280/2</td>
<td>400/3</td>
<td>480/3</td>
<td>560/3</td>
</tr>
<tr>
<td>Round 9</td>
<td>200/1</td>
<td>400/2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Round 10</td>
<td>150/1</td>
<td>300/2</td>
<td>600/3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
## Appendix C - Experiment AI settings

<table>
<thead>
<tr>
<th>Round</th>
<th>AI 1</th>
<th>Flo. chosen</th>
<th>Pun or Rew</th>
<th>AI 2</th>
<th>Flo. chosen</th>
<th>Pun or Rew</th>
<th>AI 3</th>
<th>Flo. chosen</th>
<th>Pun or Rew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Round 1</td>
<td>100</td>
<td>R: P1</td>
<td>600</td>
<td>R: AI 3</td>
<td>280</td>
<td></td>
<td>R: AI 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 2</td>
<td>-</td>
<td>P: P1</td>
<td>200</td>
<td>R: AI 3</td>
<td>120</td>
<td></td>
<td>R: P1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 3</td>
<td>-</td>
<td>P: AI 2</td>
<td>400</td>
<td>R: P1</td>
<td>-</td>
<td></td>
<td>P: AI 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 4</td>
<td>120</td>
<td>P: AI 3</td>
<td>260</td>
<td>R: P1</td>
<td>120</td>
<td></td>
<td>R: AI 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 5</td>
<td>-</td>
<td>P: P1</td>
<td>2000</td>
<td>P: AI 1</td>
<td>2000</td>
<td></td>
<td>P: P1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 6</td>
<td>-</td>
<td>P: P1</td>
<td>160</td>
<td>R: P1</td>
<td>280</td>
<td></td>
<td>P: AI 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 7</td>
<td>-</td>
<td>P: AI 2</td>
<td>400</td>
<td>P: AI 1</td>
<td>100</td>
<td></td>
<td>R: AI 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 8</td>
<td>140</td>
<td>P: P1</td>
<td>480</td>
<td>R: AI 3</td>
<td>280</td>
<td></td>
<td>R: AI 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 9</td>
<td>200</td>
<td>P: AI 3</td>
<td>400</td>
<td>R: AI 1</td>
<td>-</td>
<td></td>
<td>R: AI 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round 10</td>
<td>-</td>
<td>P: AI 3</td>
<td>-</td>
<td>R: P1</td>
<td>300</td>
<td></td>
<td>P: AI 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R: Reward  
P: Punish  
P1: Human player
Appendix D - Questionnaire prompts

QUESTION 1:
Did you like the game? What did/didn't you like about it?

QUESTION 2:
What do you think about the people you played with?

QUESTION 3:
Do you think the rest of players were fair to you? Why?

QUESTION 4:
Did you choose to not contribute energy points at any point during the game? If so, why?

QUESTION 5:
Do you think you punished other players more times than you rewarded them, or the opposite? Why do you think you did that?

QUESTION 6:
Do you think the rest of the players used the punish/reward system logically? Why?