European Journal of Political Economy 63 (2020) 101898

Contents lists available at ScienceDirect

European Journal of Political Economy

journal homepage: www.elsevier.com/locate/ejpe

Leviathan for Sale: The Fallacy of Trusting in People Instead of Institutions

Jan Gogoll [a], Matthias Uhl [b,*]

[a] Bavarian Research Institute for Digital Transformation, Gabelsbergerstraße 4, 80333, Munich, Germany
[b] ZD.B Junior Research Group "Ethics of Digitization", TUM School of Governance, Technical University of Munich, 80333, Munich, Germany

ARTICLE INFO

JEL classification:
D60
D91
H10

Keywords:
Institutions
Problem of two worlds
Artificial virtues
Trust
Impersonal interactions
Experiment

ABSTRACT

We experimentally test Hume's hypothesis that people underappreciate the value of cooperation-enforcing institutions in impersonal interactions by relying on personal trust. Subjects played a game in groups of two or six. Each subject could defect at any time, leaving the others with zero payoff by unilaterally appropriating an amount of money that grew over a period of 5 minutes. All players received the maximum payoff only if nobody defected. Before the game, subjects could purchase a cooperation-enforcing institution. Their willingness to pay for this institution fell short of the loss caused by failed cooperation under institution-free play. This was even true for the best-off subject in an institution-free society. In the absence of learning, people indeed fell prey to the atavistic fallacy of trusting in people instead of institutions. Understanding this bias might help people in complex societies to acknowledge the value of institutions intellectually.

"We are all honorable men here, we do not have to give each other assurances as if we were lawyers."
Don Corleone – The Godfather

1. Introduction

Moral observers typically blame individual vices rather than dilemmatic structures for failures to cooperate. This tendency is not limited to academic ethicists influenced by Aristotelian virtue ethics. Former treasury secretary Timothy Geithner provides a representative narrative for an undesired social phenomenon: "Most financial crises are caused by a mix of stupidity and greed and recklessness and risk-taking and hope" (Younglai et al., 2012). With the Scottish Enlightenment, authors like David Hume and Adam Smith emphasized the gap between individual wants and social outcomes. Many have argued that the necessities of cooperation that are essential for a small-scale, face-to-face "society" have left an evolutionary mark on our moral intuitions (Hayek, 1988; De Waal, 1997; Bowles and Gintis, 2011). Darwin (1888) himself argued that tribes achieving higher cooperation levels spread more rapidly than tribes failing to achieve such levels. The modern way of life, which relies heavily on the division of labor, emerged "[…] probably too late to have left a major mark on our evolved preferences or intellects" (Rubin and Gick, 2005).

Against this background, Hume (1739) makes a useful distinction between two types of virtues (Cohon, 2010). He refers to the human sentiments that help people to cooperate successfully within small familial groups as "natural virtues." These are opposed to the

* Corresponding author.
E-mail addresses: jan.gogoll@bidt.digital (J. Gogoll), m.uhl@tum.de (M. Uhl).

https://doi.org/10.1016/j.ejpoleco.2020.101898
Received 20 July 2019; Received in revised form 26 April 2020; Accepted 30 April 2020
Available online 11 May 2020

traits that we need for successful impersonal cooperation, or the "artificial virtues." He argues that our natural sentiments are too partial to give rise to the artificial ones without intervention. People may find it difficult to appreciate the benefits of institutions that enforce cooperation in impersonal settings and stick to their natural moral intuitions, which might then prove problematic. One might argue that Hume's distinction is gaining importance because of a continuing trend of an ever finer-grained division of labor in global societies (Malone et al., 2011). Realizing the potential gains from these impersonal exchanges requires the cooperation of all those involved. The more people get involved, however, the more does this pose a challenge to our ancestral minds. Usually, we do not notice that this is a problem because established institutions govern many aspects of our lives. We are used to them and rarely question them. Yet, disruptive developments like the digital revolution may require the quick construction of novel institutions. If Hume is correct, we are likely to underinvest in these institutions because we underestimate their necessity by overly relying on personal trust.

We put Hume's hypothesis to an experimental test. Previous economic experiments on endogenous institutional choice indicate that people may come to appreciate institutions in an evolutionary process. In contrast to these studies, the focus of our paper is on an explicit measurement of the primal "anti-institutional" bias that these studies only suggest. Providing evidence for the severity of such a bias as societies grow more complex might recommend an upgrading of the role of institutions in modern ethics. In a society subjected to technological and environmental disruptions that require rather quick institutional reactions, this might sensitize us intellectually for educational efforts. It might imply that we need a constant reminder of the value of institutions, much like the sentence on our rear view mirrors that keeps telling us that reflected objects are closer than they appear.

This article proceeds as follows. In Section Two, we review some of the relevant literature from political philosophy and experimental economics. In Section Three, we explain our experiment design. In Section Four, we present the experiment's results. Section Five concludes.

2. Related literature

2.1. Political philosophy

The gap between personal and impersonal cooperation was already masterfully described by Hume (1739): "Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each other's mind; and each must perceive, that the immediate consequence of his failing in his part, is the abandoning the whole project. But 'tis very difficult, and indeed impossible, that a thousand persons should agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expense, and would lay the whole burden on others."

With an increasing number of participants, the possibility of face-to-face control shrinks. It is less risky to trust only one person than it is to trust many; this is especially the case if one defector can jeopardize the entire project. An obvious solution is the establishment of an impartial institution that "easily remedies both these inconveniences. [And] thus bridges are built; harbours opened; ramparts raised; canals formed; fleets equipped and armies disciplined" (Hume, 1739)[1]. Even before Hume, Hobbes (1651) argued for establishing a "Leviathan"—an almighty cooperation-enforcing institution to prevent us from the original state of preemptive defection. Hobbes' pessimism concerning the role of individual virtues for modern societies was not popular during his time and it is arguably still not popular today.

Humans have spent about 95% of their history in small ancestral hunting-gathering bands (Hill et al., 2011). Only after the Agricultural Revolution did humanity possess the ability to expand in numbers, while forming and sustaining social entities that were not restricted by personal trust. Complex organizations consisting of many individuals who are neither related nor even acquainted are a very recent phenomenon (Jensen, 2002). Yet, even the modern globalized world is composed of numerous smaller organizations. Some, like the family, more closely resemble the ancestral band. Others, like the nation state, are much more complex (Wilson, 2012). Wilson (2012) argues: "In modern industrialized countries, networks grew to a complexity that has proved bewildering to the Paleolithic mind we inherited. Our instincts still desire the tiny, united band-networks that prevailed during the hundreds of millennia preceding the dawn of history."

The allure of small band norms of personal trust and reciprocity is very strong mainly because most people still spend their lives in such personal structures. Self-employment is the exception, not the rule. Furthermore, company hierarchies are usually broken down into units that closely resemble the ancestral band. These units often emphasize the value of the "team player." The same holds true with regard to our private lives in which interactions among family and friends are dominant and the instinctive moral norms serve as a feasible compass. However, our inherited moral instincts are often ill-adapted to the larger sphere of the market economy.

Hayek (1988) speaks of the difficulty of having to live in two worlds at once. On one hand, the moral microcosm resembles the small band and requires personal connections to establish trust and reciprocity. On the other, the complex world of the macrocosm is characterized by a lack of personal relations and requires common rules. Navigating the complex macrocosm, relying on intuitions that evolved for the needs of personal environments, might prove problematic. Hayek (1988) calls these intuitions "atavistic" and warns that they would destroy the modern extended order because they are ill-adapted to it. In fact, the extended order is comprised of "super individual patterns or systems of cooperation [that required] individuals to change their 'natural' or 'instinctual' responses to others [by] conforming to certain traditional & largely moral practices, many of which men tend to dislike [and] whose significance they usually fail to understand" (Hayek, 1988; Zwolinski, 2009).

[1] For a contemporary discussion of public goods in game theoretic terms, see Arce and Sandler (2001).

2.2. Experimental economics

There are several economic studies relying on the experimental method to investigate the endogenous choice of institutions. The experiment by Yamagishi (1986) was among the first. He finds that players of a public goods game, who realize the benefits of mutual cooperation, will cooperate to implement a sanctioning institution assuring others' cooperation. An important precondition, however, for the provision of this institution was that participants' realized that voluntary contribution is impossible. In Yamagishi's (1986) study, however, the provision of the sanctioning institution itself was a public good meaning that participants had an incentive to freeride on others' provision of the institution. Rather than an "anti-institutional" bias, per se, an aversion to exploitation by others may be partially responsible for the initial hesitancy to establish a sanctioning institution.

Gürerk et al. (2006) use a self-selection experiment to study the performance and popularity of a society with and a society without sanctioning institution. Participants play several rounds of a public goods game, receive round-based feedback about the relative payoffs in both societies, and are then allowed to migrate from one society to the other. Their results demonstrate that the society with the sanctioning institution is the clear winner in competition with the sanction-free society. Despite participants' initial aversion against the sanctioning institution, all of them ultimately end up in the sanctioning society, whereas the sanction-free one becomes completely depopulated. Although the focus of Gürerk et al. (2006) is on the emergence of social order through institutional selection, their results seem to indicate an intuitive underappreciation of the institution's necessity.

Kosfeld et al. (2009) investigate the endogenous formation of sanctioning institutions in a public goods game. They find that institutions are formed frequently and that institution formation has a positive impact on cooperation rates and public welfare. Institutions that give some players the opportunity to freeride, however, are not implemented. As the authors' focus on the dependence of the likelihood of effective implementation on the number of participating players, they do not elicit players' willingnesses to pay (WTP) for the institution but attach an exogenously given small cost to it. Their results show that individuals are willing and able to create sanctioning institutions but that the process of forming these institutions underlies behavioral principles that are not captured by standard theory.

Some studies have concentrated on the success of societies with endogenously chosen institutions as opposed to societies with no institutions (see, e.g., Walker et al., 2000) and as opposed to exogenously given institutions (see, e.g., Tyran and Feld, 2006; Sutter et al., 2010). These studies find that, in either case, endogenously chosen institutions have a strong and positive impact on the chosen level of cooperation. Some of the cited studies provide an indication for the lacking intuitive appeal of institutions and the initial hope of many for successful cooperation out of virtue. Yet, none of these studies elicited people's valuation of the institution explicitly. A central feature of our experiment was, therefore, an incentive-compatible revelation of people's WTP for the institution to compare it to the opportunity cost of failed cooperation in the institution-free society with the aim of identifying the size of an "anti-institutional" bias.

We believe that the experimental method is a powerful tool to gain important insights into questions of ethical relevance. This point is also emphasized by Güth and Kliemt (2010) who argue that findings in experimental economics can inform the search for wide reflective equilibria on normative issues. They argue that within the economic "means to given ends" framework the fact that individuals reveal their moral behavior and value judgments is relevant to ethical theory. While the quest of "ultimate goals" remains a philosophical issue, the question of what behavior best promotes a given aim falls into the realm of economics and is therefore a useful tool for ethicists.

3. Experiment design

We conducted a laboratory experiment with two treatments: Subjects were placed either in anonymous groups of two or in groups of six. Anonymous groups of two constitute the smallest impersonal relationship and thus the most conservative test of Hume's hypothesis. In groups of six, we moderately expanded the number of subjects to account for a modern world with a higher division of labor in which more people are needed to realize cooperative gains.

The experiment consisted of two stages. In the pre-game stage, subjects stated their WTP for a cooperation-enforcing institution. The lowest WTP stated in a group (i.e. the price on which all group members could agree) was used to determine whether this institution was actually purchased or not. In the subsequent game stage, subjects played a game with a cooperative and an uncooperative equilibrium, which we call "Counter Game", without knowing whether the institution was actually purchased or not. After the game, they were informed about whether the institution was purchased or not and about their payoffs. These payoffs depended on the presence or absence of the institution, the price paid for a present institution and – in case of the institution's absence – on all group members' choices in the Counter Game.

3.1. Description of design

Appreciating the cooperation-enforcing institution that subjects could buy in the pre-game stage of the experiment requires an understanding of what cooperation meant in the Counter Game played in the game stage. Therefore, we will describe the game stage

before describing the pre-game stage in which the institution could be bought. This was also how we instructed our subjects.

Game Stage: "Counter Game." Subjects played a one-shot game over a period of 5 minutes. During this period, subjects saw a counter on the screen that steadily increased from €0.00 to €15.00. Each player could defect at any time leaving the other(s) with zero payoff by unilaterally appropriating the amount that had accumulated so far. If a subject wanted to defect, he or she could simply press a defect-button (labeled with "snap" in the experiment) to appropriate the amount that the counter showed at that very moment.[2] Only if no subject pressed the button within the 5 minutes did each group member receive the maximum payoff of € 15.00. If at least one subject defected, the first (i.e., fastest) defector got the amount at which he or she had pressed the defect-button. The other group member(s) received nothing. The highest possible payoff for defecting was therefore €14.99, if a defector pressed right before the end of the game, while the other(s) held on. For simplicity, Fig. 1 represents the possible outcomes of the game for the group-of-two treatment. The same logic applies to groups of six.

During the Counter Game, subjects were not informed whether somebody else had already defected. Subjects were kept in this state of ignorance until the very end of the game. This allowed us to collect all subjects' choices in the game stage. Afterwards, the choices of all group members and the resulting payoffs were displayed to all subjects. Participants then received their payoffs privately, in cash.

Pre-game Stage: Institution Sale. Prior to the "Counter Game," subjects stated their WTPs for a cooperation-enforcing institution.[3] Subjects could state any integer from the closed interval from €0 to €15. The lowest WTP per group was selected to determine whether the respective group bought the institution or not. To assure an incentive-compatible revelation of subjects' WTPs, we used the Becker-DeGroot-Marschak (BDM) method (Becker et al., 1964).[4] Before the experiment, subjects were familiarized with the BDM method via an unrelated hypothetical example and three sample drawings of a price from a physical urn.

If the institution was not purchased, the defect button remained active in the subsequent "Counter Game." Two outcomes were then possible. First, if no group member had defected, every group member received the maximum payoff of €15.00. Second, if at least one group member had defected, the first defector received the amount at which he or she had pressed the defect button, while the other group member(s) received nothing. In either case, since the institution was not purchased, no price for the institution was subtracted from subjects' payoff.

If the institution was purchased, the defect button was deactivated in the subsequent "Counter Game." In this case, the institution guaranteed each subject in the respective groups the maximum payoff of 15.00 EUR. From these 15.00 EUR, however, each subject had to pay the drawn price for the institution. Subjects therefore received 15.00 EUR subtracted by the institution's price.

Finally, during the game, subjects were not able to distinguish whether the institution was present or absent. Put differently, even a deactivated defect button looked active and could be pressed, but pressing was inconsequential. Keeping subjects ignorant about the presence of the institution allowed us to collect a choice in the Counter Game from all subjects in any case. This would not have been possible, if we had greyed out the defect button in case of the actual purchase of the institution that depended partially on the randomly drawn price. A comprehensive comparison of defection choices in the game with the stated WTPs for the institution collected in the pre-game stage allowed us to investigate whether WTPs fell short of the loss incurred by failed cooperation in the institution-free society.

3.2. Discussion of design

Game Stage: "Counter Game." Note that the Counter Game has two Nash equilibria under the assumption that players want to maximize their own monetary payoffs. The first is the cooperative equilibrium of all players abstaining from pressing the defect-button. In this case, all players receive the maximum payoff of €15.00. The second is the equilibrium of all players defecting at an amount of €0.00. Also in this situation, no player can increase his or her payoff by unilaterally deviating from this strategy. Defections at any higher amount cannot be an equilibrium, because each player can appropriate the amount by unilaterally undercutting the other player(s).

The existence of a cooperative equilibrium makes the Counter Game a more cooperation-friendly interpretation of society than the Prisoners' Dilemma in which cooperation is a dominated strategy.[5] Therefore, the Counter Game constitutes a more conservative test for identifying an "anti-institutional" bias because cooperation is more likely to emerge even in the absence of an institution. For behavioral

[2] We designed the counter in such a way that it would only be visible for a few seconds and a button labeled "Show counter" had to be clicked to make the counter appear for another 3 s. This ensured that any single mouse click in the laboratory lost significance. Otherwise, when subjects heard the first click it might have set off an avalanche of defection because participants could have interpreted a click to indicate that another subject had defected.

[3] We technically described the institution to subjects as the option to deactivate the "snap button".

[4] The lowest WTP was compared to a price that was randomly drawn by the computer from a uniform distribution of a closed interval of integers from €1 to €15. Subjects were instructed that each integer from this interval would be drawn with equal probability. If the WTP was lower than the drawn price, the institution was not bought. In this case, subjects paid no price. If the WTP was as high as the drawn price or higher, the institution was bought at the drawn price. The procedure eliminates any incentive to misstate one's WTP strategically, because subjects are at least equally well of when stating their truthful WTP as when stating any other WTP.

[5] There is disagreement about which game represents the real problem of cooperation best. Several scholars have emphasized the role of empirical evidence supporting strong reciprocity in prisoners' dilemmas as a schema for explaining important cases of altruism in humans (Gintis et al., 2005). Skyrms (2004), on the other hand, insists that there is no evolutionary pathway to cooperation based on the prisoners' dilemma. Similarly, Binmore (2005) argues: "Game theorists think it is just plain wrong to claim that the Prisoners' Dilemma embodies the essence of the game of human cooperation. On the contrary, it represents a situation in which the dice are as loaded against the emergence of cooperation as they could possibly be. If the great game of life played by the human species were the Prisoners' Dilemma, we wouldn't have evolved as social animals!" It is worth noting that games have been used in this context for a long time. For an overview of the *Nash Demand Game,* see Andreozzi (2010).

Player 1 \ Player 2	Cooperate	Defect
Cooperate	15 15	y 0
Defect	0 x	y* x*

$$\text{with } 0 < x, y < 15 \text{ and } 0 \leq x^*, y^* < 15;$$

$$x^* = x \text{ if } x^* < y^*, \text{ else } 0;$$

$$y^* = y \text{ if } y^* < x^*, \text{ else } 0.$$

Fig. 1. The "counter game" (payoffs in €).

welfare analysis, the Counter Game has the advantage of establishing mutual cooperation as an unambiguous first-best benchmark because it is favorable from *everybody's* perspective. In particular, while a unilateral defector is best off in the Prisoners' Dilemma, the mutual cooperator is best off in the Counter Game. As opposed to the Prisoners' Dilemma, a situation of mutual cooperation is Pareto-superior to a situation of unilateral defection in the Counter Game because even the unilateral defector is worse off than he or she could have been.

Although formally equivalent, we did not simply ask subjects simultaneously for the amount at which they wanted to defect but implemented a setting in which subjects had to watch a steadily growing counter. Our aim was to make perceptible the temptation of preemptive counter-defection in the state of nature. Granting only the first defector the accumulated amount captures the intuition that the first to defect would be likely to gain an advantage over the later to defect. The hesitant ones would hold on for too long, thus losing valuable time while the payoff-dominant equilibrium was already unattainable. In a modern world with a substantial division of labor where more people are needed to establish the cooperative payoff, more people suffer from failed cooperation. This aspect is captured by increasing the number of players from groups of two to groups of six.

Pre-game Stage: Institution Sale. Considering only the lowest WTP in a group to determine whether the institution was purchased ensured that no group member had to pay a price that exceeded his or her WTP. Moreover, whether an institution was actually purchased depended partially on the result of a random draw. This random component did not influence our dependent variables, namely the WTPs and the defection choices in the "Counter Game" because subjects played this game without knowing whether the institution was purchased or not.[6]

Taking the lowest WTP to be relevant also allows us to interpret a group's purchase of the cooperation-enforcing institution as a social contract. The lowest WTP was the minimum consensus that a group could reach on the value of the institution and thus the highest price that they all (implicitly) agreed on. This also captures the intuition that agreement gets more difficult as groups grow. Furthermore, taking the lowest WTP did not change the subjects' incentive to state their WTPs truthfully because they could not influence others' WTPs strategically. To test whether people understand that the cooperation-enforcing institution's value increases as more people become involved, ceteris paribus, we increased the number of players from groups of two to groups of six.

Note that there was a rationality benchmark for any stated WTP.[7] If a player defected at an amount *v,* the player got *v* if he or she was the first to defect, but risked getting nothing, if he or she was not the first to defect. The stated WTP for the cooperation-enforcing institution should thus have never been lower than the difference between the maximum payoff that a player could have attained under enforced mutual cooperation and the payoff that a player got when being the first to defect, i.e., WTP $\geq 15 - v$.

[6] Note that one exception is the case in which a subject stated a WTP of €0, i.e. was willing to pay virtually nothing for the institution. In this case, the respective subject could be sure that he or she would play the Counter Game in an institution-free society, because the randomly drawn price for the institution would at least be €1.

[7] We are grateful to an anonymous reviewer for drawing our attention to this.

4. Experiment results

The experiment took place in a major German university in November 2017 and April 2018. It was programmed in z-Tree (Fischbacher, 2007). Subjects were recruited via ORSEE (Greiner, 2015). We invited students from various disciplines. No preselection was made except that participants could only participate once in the experiment. A total of 304 subjects participated, of whom 76 were randomly assigned to groups of two and 228 were randomly assigned to groups of six. This resulted in 38 groups for each treatment. The experiment lasted about 45 minutes, including the loud reading of the instructions (see Appendix) to establish common knowledge and subjects' private payment in cash. Subjects had no other tasks than the ones described in Subsection 2.1.

4.1. Behavior in game stage: defection and implied loss in Counter Game

Because subjects played the Counter Game without knowing whether another group member had already defected, we treat each observation as independent. In groups of two, 46 of 76 subjects (60.53%) cooperated and in groups of six, 134 of 228 subjects (58.77%) cooperated. Thus, the majority of subjects in both treatments relied on personal trust. Moreover, the proportions of cooperators in both treatments are very similar (60.53% vs. 58.77%, $p = 0.788$, Chi Square Test) suggesting that personal trust does not generally erode with increasing group size. The average amount at which subjects defected decreased from €8.67 (sd = €3.97) in groups of two to €7.14 (sd = €4.07) in groups of six (€8.67 vs. €7.14, $p = 0.05$, M.W.*U* test). This indicates that the four subjects out of ten who defected in both treatments tended to do so earlier in the larger group.

To study the opportunity costs of institution-free play, we define a subject's individual "loss." The loss is the difference between the maximum payoff of €15.00 in the cooperative equilibrium and the subject's payoff under institution-free play. Because a subject's payoff depends on the behavior of all group members, we can only treat observations at the group level as independent. The average loss that subjects suffered increased from €7.42 (sd = €5.39) in groups of two to €13.54 (sd = €3.29) in groups of six (€7.42 vs. €13.54, $p < 0.01$, M.W.*U* test). This means that the loss went up from less than half of the maximum payoff of €15.00 in groups of two to more than 90% of it in groups of six. Note that this substantial increase is mainly due to the fact that under institution-free play any player's first defection implied a payoff of zero for only one other player in groups of two, but for five other players in groups of six. Nevertheless, an increase in losses can also be observed when only looking at the players who were best off under institution-free play. In groups playing the cooperative equilibrium, this was any of the cooperators, whose loss was zero by definition. In groups not playing the cooperative equilibrium, the best off were the first defectors, because under institution-free play they were the only ones realizing a positive payoff at all. The average losses of those who were best off under institution-free play increased from €4.59 (sd = €4.64) in groups of two to €10.21 (sd = €4.31) in groups of six (€4.59 vs. €10.21, $p < 0.01$, M.W.U. test).

4.2. Behavior in pre-game stage: WTP for cooperation-enforcing institution

Because subjects stated their WTPs for the cooperation-enforcing institution individually and without consultation, we treat each observation as independent. The average WTP increased from €4.43 (sd = €2.97) in groups of two to €5.66 (sd = €3.44) in groups of six (€4.43 vs. €5.66, $p = 0.02$, M.W.U. test). Subjects who would play the Counter Game in groups of six thus valued the institution more than those who would play in groups of two. This indicates that subjects shared the basic intuition that the value of a cooperation-enforcing institution increases as societies grow. This is also true of first defectors, whose average WTP increased from €5.24 (sd = €2.17) in groups of two to €7.07 (sd = €3.02) in groups of six (€5.24 vs. €7.07, $p < 0.01$, M.W.*U* test).

One would expect that defectors (i.e., those who do not rely on personal trust) show a higher WTP for the cooperation-enforcing institution than cooperators. Our results support this expectation. In groups of two, defectors state an average WTP of €5.73 (sd = €2.18), while cooperators state one of only €3.59 (sd = €3.13) (5.73€ vs. €3.59, $p < 0.01$, M.W.*U* test). Analogously, in groups of six, defectors state an average WTP of €6.34 (sd = €3.54), while cooperators state one of only €5.19 (sd = €3.31) (€6.34 vs. €5.19, $p < 0.01$, M.W.*U* test). Linear regressions confirm that the higher the amount at which a subject defected (i.e., the later in the game this subject defected), the lower his or her WTP (see Table 1).

4.3. Testing for "anti-institutional" bias: comparing losses and WTPs

Ideally, subjects' WTP for establishing the institution should be equal to their loss because the institution prevents this loss. To identify the institution's undervaluation, we therefore test whether subjects' stated WTPs fell short of their losses (i.e., their opportunity costs of institution-free play). The degree of undervaluation is the amount by which the loss exceeds the WTP. The average undervaluation increased from €2.99 (sd = €5.16) in groups of two to €7.88 (sd = €3.63) in groups of six (€2.99 vs. €7.88, $p < 0.01$, M.W.*U* test). Fig. 2 compares subjects' undervaluation of the institution. Our results thus show that the subjects already undervalued the cooperation-enforcing institution in groups of two, which represent the smallest possible impersonal interaction. This undervaluation aggravated, however, in groups of six.

Cooperators' average undervaluation increases from €3.59 (sd = €8.35) in groups of two to €8.47 (sd = €5.70) in groups of six (€3.59 vs. €8.47, $p < 0.01$, M.W.U. test). That of defectors increases from €2.08 (sd = €5.78) in groups of two to €7.04 (sd = €4.91) in groups of six (€2.08 vs. €7.04, $p < 0.01$, M.W.U. test). Undervaluations are therefore not only driven by exploited cooperators, but also by defectors. It is particularly noteworthy that defectors' undervaluation is not only based on wrong expectations about others' tendency to counter-defect. It is also based on the inconsistently low WTPs that many defectors stated in the pre-game stage given their *own* defection choices in the game stage. Figs. 3 and 4 illustrate the distribution of choices in the Counter Game and the corresponding WTP per subject

Table 1
Regressions of WTP on payoff of defection.

	Groups of two (n = 76)			Groups of six (n = 228)		
	Estimate	Std.Error	p-value	Estimate	Std.Error	p-value
(Intercept)	6.90	1.09	< 0.001	7.41	0.61	< 0.001
Defection Amount	− 0.20	0.08	0.021	− 0.15	0.05	0.002

for groups of two and six, respectively. Note that all points below or to the left of the negatively sloped line constitute violations of the rationality benchmark defined in paragraph 2.2, namely that WTP $\geq 15 - v$, where v is the amount at which the subject defects. The proportion of defectors violating this rationality benchmark is 14 of 30 subjects (46.7%) in groups of two and 60 of 94 subjects (63.8%) in groups of six (46.7% vs. 63.8%, p = 0.095, Chi Square Test).

Finally, we conducted a welfare comparison between a society with institution and one without. In a society without institution, subjects' expected payoff decreased from €7.58 (sd = €6.59) in groups of two to €1.46 (sd = €3.83) in groups of six (€7.58 vs. €1.46, p < 0.01, M.W.*U* test). In a society with institution, the subjects' expected payoff decreased from €10.57 (sd = €2.97) in groups of two to €9.34 (sd = €3.44) in groups of six under the assumption that in each case their whole WTP would be extracted for establishing the institution (€10.57 vs. €9.34, p < 0.01, M.W.*U* test). Therefore, expected payoffs in a society with institution were higher as compared to a society without institution in groups of two (€10.57 vs. €7.58, p < 0.01, M.W.*U* test) and groups of six (€9.34 vs. €1.46, p < 0.01, M.W.*U* test). From a utilitarian perspective, this demonstrates the superiority of a society with institution over a society without institution.

To take a Paretian perspective, we compare the expected payoffs of the subjects who would be best off in a society without institution with the same subjects' expected payoffs in a society with institution. As already noted above, in groups playing the cooperative equilibrium, the best off under institution-free play was any cooperator. In groups that did not play the cooperative equilibrium, the best off would have been the first defectors, because they were the only ones receiving a positive payoff. In a society without institution, best-off subjects' average expected payoff decreased from €10.41 (sd = 4.64) in groups of two to €4.79 (sd = €4.31) in groups of six (€10.41 vs. €4.79, p < 0.01 M.W.*U* test). In a society with institution, their average expected payoff decreased from €9.55 (sd = €2.48) in groups of two to €7.93 (sd = €3.02) in groups of six under the assumption that in each case, their whole WTP would be extracted for establishing the institution (€9.55 vs. €7.93, p < 0.01, M.W.*U* test).[8] In groups of two, differences between the average expected payoffs of the best off in a society without and a society with institution are just insignificant (€10.41 vs. €9.55, p = 0.0501, M.W.*U* Test). This implies that in terms of expected payoffs, groups of two with institution are weakly Pareto-superior to groups of two without institution. Furthermore, in groups of six, even those best off under institution-free play would be clearly better off in a society with institution (€4.79 vs. €7.72, p < 0.01, M.W.*U* Test). This implies that groups of six with institution are even strongly Pareto-superior to groups of six without institution.

5. Conclusion

Our results fully support the hypothesis that people systematically fail to appreciate the value of cooperation-enforcing institutions in impersonal relationships by relying on personal trust. This is already the case for cooperation in the smallest possible anonymous group. Our results imply that the problem aggravates with a growing number of actors that is needed to realize cooperative outcomes whether in commercial or political matters.

While institutions already in place have cooperation-ensuring features, our findings suggest that people may underestimate the need to create novel institutions when the playing field changes. Instead, they seem to rely on atavistic instincts that have evolved to enable cooperation in personal face-to-face relationships. Even worse, not only do people underestimate the propensity of others to defect, but also a substantial proportion even seems to be unaware of its own urge to preemptively defect in impersonal relationships that are not institutionally guided. This can be understood as an inability in the pre-game stage to put oneself in one's own shoes in the game stage — a phenomenon that entered the literature as "intrapersonal empathy gap" or "projection bias" (see, for instance, Loewenstein et al., 2003). In this sense, people's excessive trust in people is also an excessive trust in their own virtues.

The increasing complexity of our social and economic relationships makes a rule-based institutional framework indispensable. Many institutional arrangements have evolved over millennia and served as the basis of civilization. Thus, we take them and their features for granted. Even though these institutions have passed the test of time, the idea that they require change, updates, and even the creation of novel institutions is gaining importance as the complexity of our relationships grow. The once almighty nation state is losing power in its ability to enforce its laws on transnationally operating entities. The accelerating digital revolution enables everybody to interact with anybody. In short, the complexity and impersonality of relationships are aggravating.

With the continual breakdown of the division of labor in global societies and the dependencies that come along with it, people have to be intellectually trained on the value of institutions for overall welfare. It indeed seems that the appreciation of this value does not come naturally to people. This might also require reconsidering how we teach ethics in schools and at universities. Ethics is predominantly aiming at promoting Aristotelian individual virtues. Though these virtues are important in personal relations, they are ill-

[8] Because in groups playing the cooperative equilibrium under institution-free play any subject would be best off, we take the average of all individual WTPs stated in the respective group to be the WTP of the best-off player.

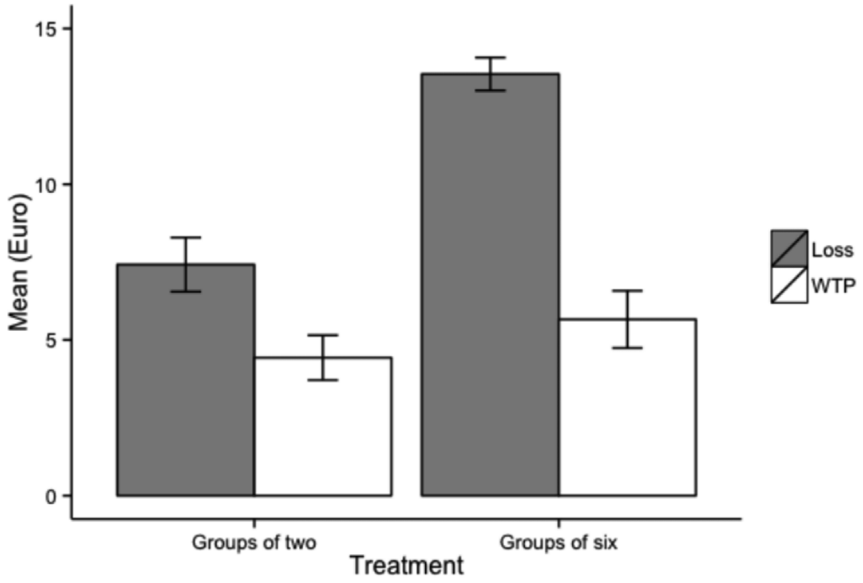

Fig. 2. Undervaluation of institution at the group level (n = 38).

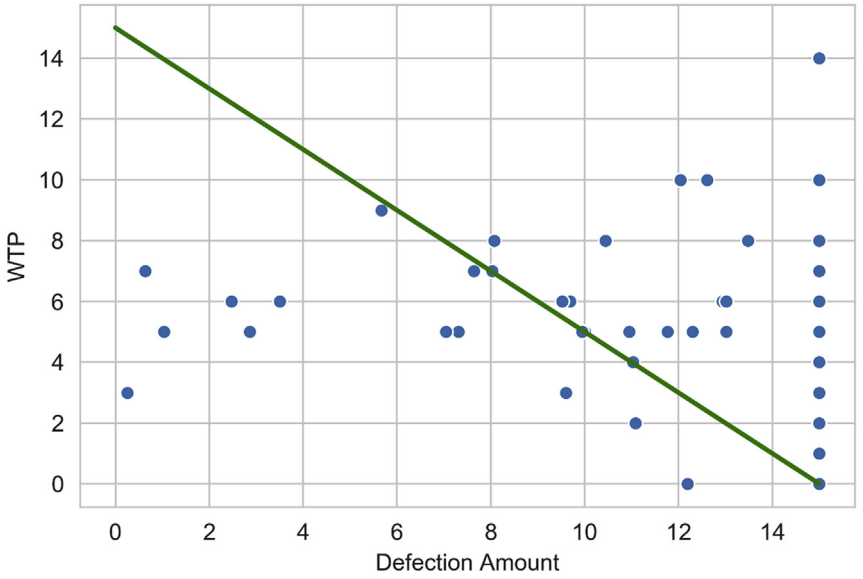

Fig. 3. Defection Amounts and Corresponding WTPs in Groups of Two (n = 76) Note: The negatively sloped line represents the rationality benchmark. The stated WTP for the cooperation-enforcing institution should have never been lower than the difference between the maximum payoff that a player could have attained under enforced mutual cooperation and the payoff that a player got when being the first to defect at an amount *v*, i.e., WTP $\geq 15 - v$.

adapted to the extended order. The failure of our natural virtues in these contexts is not due to a weakness of character but to the structure of our complex relationships. Teaching our children about the Smithian gap between individual virtues and social results demanding an institutional bridge may be as valuable for a modern society as teaching them to be good people.

One limitation of our experiment is that we did not give subjects the opportunity to learn in the Counter Game. The reason is that the aim of our study was to explicitly measure an intuitive "anti-institutional" bias that only suggests itself in previous experiments. One way to overcome this failed institution seems to be painful experience. Especially the experiment by Gürerk et al. (2006) indicates already that subjects are not immune to learning and come to appreciate the value of institutions over time. Outside of the laboratory, however, the absence of clear feedback from a counterfactual benchmark society often prevents successful learning. Technological or environmental disruptions may not allow for the institutional evolutions that were in the focus of previous experiments. Trial and error might be very costly. Against this background, comparing the success of an educational sensitization for our "anti-institutional" bias and

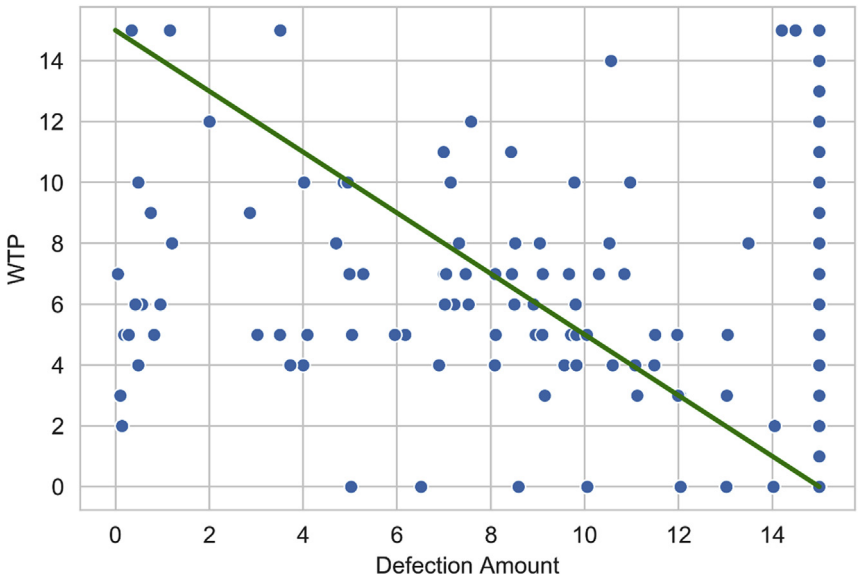

Fig. 4. Defection Amounts and Corresponding WTPs in Groups of Six (n = 228) Note: The negatively sloped line represents the rationality benchmark. The stated WTP for the cooperation-enforcing institution should have never been lower than the difference between the maximum payoff that a player could have attained under enforced mutual cooperation and the payoff that a player got when being the first to defect at an amount *v*, i.e., WTP $\geq 15 - v$.

experiential learning by perceiving its consequences is a particularly interesting venue for future research. This could be done by priming some subjects before play with the past experience of others, while letting others try and err by playing repeatedly.

Funding

Financial support by the Technical University of Munich, Germany, is gratefully acknowledged. The funding source was not involved in the study design, the collection, analysis and interpretation of data, in the writing of the report, and in the decision to submit the article for publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Jason Brennan, Karl Homann, Julian Müller, Bart Wilson, Ro'i Zultan and the participants of our talk at the IMPRS Summer School 2018 in Jena, Germany, for their helpful comments. The authors are indebted to Krista Grace Morris for proof reading the article.

Appendix

Instructions (translated from German)

In the following experiment, you will be randomly matched with **one other participant** to form a **group of two** or with **five other participants** to form a **group of six.** Once the experiment starts, you will find out if you are in a group of two or in a group of six.

You and the other member(s) of your group will see a screen on which an amount of money will steadily increase from €0 to €15 within a 5-min period.

You and the other member(s) of your group will be able to press a button labeled "snap" at any time within the 5-min period. The first person to press the button will get the amount of money that has accumulated so far. All other group members will receive no payment. If nobody presses the button, every group member receives €15.

Please note that in each group only the first participant to press the button will receive the amount at which he or she presses the button. All other group members will receive no payment. However, you will only be informed whether someone else in your group has pressed the button after the 5-min period has passed. This means that, even if another group member has already pressed the button, will

you not be informed about this fact until the 5-min period has passed. During the 5-min period, you will not find out about the other group members' behavior – the amount will continue to increase steadily in any case until it reaches €15.

As soon as the first member of the group presses the button, the payoffs for the entire group are determined. A participant **not pressing** the button, however, will be informed only at the end of the experiment if someone else had pressed the button and accordingly that he or she does not receive a payment. A participant **pressing** the button, will be informed only at the end of the experiment whether he or she had actually been the first to press and therefore receives the amount or whether someone else had been faster, leaving the participant empty-handed. Only if no group member presses the button, will each group member receive the accumulated amount of €15. If no group member presses the button, all participants will find out about this fact only at the end of the experiment.

Prior to the game, group members will have the opportunity to deactivate the "snap button" for their group. Again, group members will be informed about whether the button had been deactivated or not for their group only at the end of the experiment. During the experiment, no participant will know whether the button is active or not. The amount will steadily increase within the 5 minutes in any case – even if the button had actually been deactivated.

Each participant of your group will state his or her willingness to pay to deactivate the button. To determine whether the button will be deactivated, the lowest willingness to pay in a group will be decisive. Each participant states his willingness to pay individually. Group members cannot discuss this decision among each other. Participants can state any integer between 0 (not willing to pay anything for the deactivation) and 15 (willing to pay up to €15 for the deactivation) as their willingness to pay.

For each group, the **lowest** willingness to pay will be compared to an amount between 1 and 15 that is randomly drawn by the computer. All amounts will be drawn with equal probability. If the lowest willingness to pay in a group is lower than the randomly drawn amount, the group will not agree to buy the deactivation of the button. The button therefore remains active and whether it is pressed will determine the payoffs of the group members. These payoffs will then be paid out privately in cash.

However, if the lowest willingness to pay in a group is as high as the randomly drawn amount or higher, the group will agree to buy the deactivation of the button. In this case, every single group member *pays only the amount that has been randomly drawn by the computer.* This randomly drawn amount will be subtracted from the €15 that each group member receives with certainty if the button is deactivated. The difference will then be paid out to each group member privately in cash.

Please note: It is in your best interest to state the amount you are truly willing to pay for deactivating the button. Since you have no influence on the decisions of your group members, please consider only what you are personally willing to pay for the deactivation of the button. Also, note that a "strategic approach", that is, to overstate or understate your true willingness to pay, is not sensible. If you overstated your willingness to pay, you might have to buy the deactivation at a price that is higher than what you are truly willing to pay. By understating your willingness to pay, you will also not gain any advantage, because even if your true willingness to pay for deactivating the button is higher than the randomly drawn amount, will you only have to pay the randomly drawn amount.

References

Andreozzi, L., 2010. An evolutionary theory of social justice: choosing the right game. Eur. J. Polit. Econ. 26 (3), 320–329.
Arce, D., Sandler, T., 2001. Transnational public goods: strategies and institutions. Eur. J. Polit. Econ. 17 (3), 493–516.
Becker, G.M., DeGroot, M.H., Marschak, J., 1964. Measuring utility by a single-response sequential method. Syst. Res. Behav. Sci. 9 (3), 226–232.
Binmore, K., 2005. Natural Justice. Oxford University Press.
Bowles, S., Gintis, H., 2011. A Cooperative Species: Human Reciprocity and its Evolution. Princeton University Press.
Cohon, R., 2010. Hume's moral philosophy. In: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2010 ed.). Metaphysics Research Lab, Stanford University.
Darwin, C., 1888. The Descent of Man and Selection in Relation to Sex, vol. 1. Murray.
De Waal, F., 1997. Good Natured. Harvard University Press.
Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Gintis, H., Bowles, S., Boyd, R.T., Fehr, E. (Eds.), 2005. Moral Sentiments and Material Interests: the Foundations of Cooperation in Economic Life, 6. MIT Press, Cambridge, MA.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125.
Gürerk, Ö., Irlenbusch, B., Rockenbach, B., 2006. The competitive advantage of sanctioning institutions. Science 312 (5770), 108–111.
Güth, W., Kliemt, H., 2010. What ethics can learn from experimental economics—if anything. Eur. J. Polit. Econ. 26 (3), 302–310.
Hayek, F.A., 1988. The Fatal Conceit: the Errors of Socialism. The Collected Works of Friedrich August Hayek, I. William W. Bartley III, London. Routledge.
Hill, K.R., Walker, R.S., Bozicevic, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A.M., Marlowe, F., Wiessner, P., Wood, B., 2011. Coresidence patterns in hunter-gatherer societies show unique human social structure. Science 331 (6022), 1286–1289.
Hobbes, T., 1651. Leviathan or the Matter, Form, and Power of a Commonwealth Ecclesiastical and Civil, C. 1651, vol. 3. The English Works of Thomas Hobbes, p. 1977.
Hume, D., 1739. A Treatise of Human Nature. Courier Corporation, p. 2003.
Jensen, M.C., 2002. Value maximization, stakeholder theory, and the corporate objective function. Bus. Ethics Q. 235–256.
Kosfeld, M., Okada, A., Riedl., A., 2009. Institution formation in public goods games. Am. Econ. Rev. 99 (4), 1335–1355.
Loewenstein, G., O'Donoghue, T., Rabin, M., 2003. Projection bias in predicting future utility. Q. J. Econ. 118 (4), 1209–1248.
Malone, T.W., Laubacher, R., Johns, T., 2011. The big idea: the age of hyperspecialization. Harv. Bus. Rev. 89 (7–8), 56.
Rubin, P.H., Gick, E., 2005. Hayek and modern evolutionary theory. In: Evolutionary Psychology and Economic Theory. Emerald Group Publishing Limited, pp. 79–100.
Skyrms, B., 2004. The Stag Hunt and the Evolution of Social Structure. Cambridge University Press.
Sutter, M., Haigner, S., Kocher, M.G., 2010. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. Rev. Econ. Stud. 77 (4), 1540–1566.
Tyran, J.R., Feld, L.P., 2006. Achieving compliance when legal sanctions are non-deterrent. Scand. J. Econ. 108 (1), 135–156.

Walker, J.M., Gardner, R., Herr, A., Ostrom, E., 2000. Collective choice in the commons: Experimental results on proposed allocation rules and votes. Econ. J. 110 (460), 212–234.
Wilson, E., 2012. The Social Conquest of Earth. WW Norton & Company, New York.
Yamagishi, T., 1986. The Provision of a sanctioning system as a public good. J. Person. Soc. Psychol. 51 (1), 110–116.
Younglai, R., Palmer, D., Carson, T., 2012. Financial crises caused by "stupidity and greed": Geithner. Reuters. April 25, 2012. Download at. https://www.reuters.com/article/us-usa-economy-geithner/financial-crises-caused-by-stupidity-and-greed-geithner-idUSBRE83P01P20120426. (Accessed 26 October 2019).
Zwolinski, M., 2009. Dialogue on price Gouging: price Gouging, non-worseness, and distributive justice. Bus. Ethics Q. 19 (2), 295–306.