

Experimental Computational Philosophy: shedding new lights on (old) philosophical debates

Vincent Wiegel and Jan van den Berg¹

Abstract. Philosophy can benefit from experiments performed in a laboratory for philosophical experimentation (SophoLab). To illustrate the power of Experimental Computational Philosophy, we set up and ran several experiments on a part of Harsanyi's theory on utilitarianism. During decomposition and translation of this theory in the experimental setting of SophoLab, we discovered that it is underspecified. We filled in some blank spots and found out that information and its costs are key in the effectiveness of act and rule utilitarianism. We also identified three further elements that have particular influence on the effectiveness of both strands of utilitarianism: group size of agents, decision-making around uncertainty, and social culture towards particular types of actions. We conclude having shown that setting up computational philosophical experiments is a useful way to gain new and deeper insights in existing argumentations used in old (and new) philosophical debates.

1 Introduction

In philosophy it can be hard to test a theory in practice, i.e., on humans because it would be unethical to expose them to harm or impossible because the number of different settings is simply too large to cover all of them in test situations. In addition, it is often required to analyze a philosophical theory with respect to aspects like consistency, completeness, and soundness of reasoning. These observations invite for the creation of a laboratory where experiments can be set up to test philosophical theories.

As with all experiments, philosophical experiments (should) make use of an environment in which situations of study are reconstructed in a way that abstracts from non-relevant factors. This can be achieved by transforming the theory under examination into a different conceptual structure that exhibits the necessary features of abstraction and control. The new conceptual structure may be constructed by making use of a conceptual framework together with a set of techniques. The role of the conceptual framework is to provide a consistent set of concepts to rephrase the theory. Game theory [5] may be considered as such a conceptual framework, another one is the belief-desire-intention model [1]. Key characteristic of such a framework is that theories that have been formulated in terms of its concepts can easily be prepared for experimental testing by making use of the corresponding techniques. A computer with some software can offer such techniques creating an experimental environment for Computational Philosophy: SophoLab [6]. SophoLab is the name by which the whole of methods, techniques and systems as mentioned above, is designated. The experiments detailed below have been executed using SophoLab.

Several people have worked as a Computational Philosopher. Danielson [2, 3], for example, has constructed computer programs that represent players at a game. He uses the game and the strategies of the players that represent particular moral philosophical stances, to test these positions. In similar ways, computers are increasingly used by social scientists and philosophers to support their research activities. The study of utilitarianism can also benefit from these new means of experimental research because many assumptions and axioms of the theory can be cast in logical and mathematical forms. This motivated us to use the theory of utilitarianism provided by Harsanyi [4] as a test case for experimental computational philosophy. Therefore, the goal of this paper is to introduce the basic ideas underlying experimental computational philosophy and to illustrate the approach by analyzing experimentally Harsanyi's theory of utilitarianism.

The rest of this paper is structured as follows. In section 2 we provide a quick overview of the methodology of experimental computational philosophy. Section 3 prepares the experiments by analyzing Harsanyi's theory of utilitarianism showing some white spots. Section 4 describes the setup and running of the experiments according to the methodology described in section 2. This description includes the report of the results found. In the final section 5 we present some conclusions.

2 Experimenting in philosophy

What does it mean to run philosophical experiments? The answer to this question can not be written down in a few statements. Elsewhere, the methodology and the translation are described in more detail [6]. For the current purpose we give a short description of the steps taken in the setting up of experiments, running them, and translating back the results of the experiments. These steps are

1. Decomposing the selected philosophical theory into assumptions, premises, predictions, etc.
2. that can be translated into a concrete experimental setting,
3. and translated into the elements of the intermediate conceptual framework.
4. The concrete experimental setting must be reflected in the intermediate conceptual framework.
5. The theory is implemented in the laboratory based on the requirements of the conceptual framework,
6. reflecting the concrete experimental setting.
7. Experiments are conducted
8. and the results are translated back into the (restated) terms of the theory
9. that can be used to confirm, refine, reject, etc. the theory.

¹ Faculty of Technology, Policy, and Management, Delft University of Technology, The Netherlands, email: {v.wiegel,j.vandenbergh}@tudelft.nl

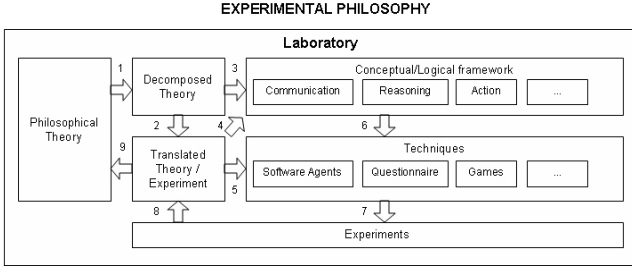


Figure 1. Steps in experimental computational philosophy [6].

As has been announced above, we illustrate the above-given general outline of experimental computational philosophy by analyzing and implementing Harsanyi's theory of utilitarianism. We first start by explaining the relevant basic issues of this theory.

3 Harsanyi's theory of utilitarianism

Harsanyi's theory of utilitarianism [4] has its roots in the work of Adam Smith, Immanuel Kant and the utilitarian tradition of Jeremy Bentham and John Stuart Mill. From Bentham and Mill he takes the concept of maximization of social utility (social welfare function) as the criterion of the morally good. He prefers rule utilitarianism over act utilitarianism and formulates his moral criterion as follows ([4], page 41):

"...a correct moral rule is that particular behavioral rule that would maximize social utility if it were followed by everybody in all situations of this particular type."

In addition, rationality plays an important role in his theory. In his view ethics is a part of a general theory of rational behavior on par with decision theory and game theory. The rationality criterion is expressed as Pareto optimality and the Bayesian rationality postulates.

Using utilitarianism as our starting point including its definition of what is morally good (i.e., maximum social utility), we focus on the claims that Harsanyi makes as to how best achieve the moral good. In order to be able to test his claims in an experimental setting, we interpret his recommended behavioral rules as strategies as used by (individual) agents. Then, Harsanyi's arguments can be formalized as follows. Let S_1, S_2, \dots, S_z be the strategies S_i of agents $1, 2, \dots, z$, where each strategy S_i is an element of the set of all possible strategies. The social welfare function $W(S_1, \dots, S_z)$ is the sum of all individual utilities:

$$W(S_1, \dots, S_z) = \sum_{i=1}^z U_i(), \quad (1)$$

where $U_i()$ is the utility of agent i . The welfare function $W(S_1, \dots, S_z)$ is maximized over the strategies S_1 to S_z of all agents:

$$W_{\max} = \max_{S_1 \dots S_z} W(S_1, \dots, S_z). \quad (2)$$

The utility function must adhere to the rationality requirements such as a complete pre-ordering and continuity.

In our discussion and experimentation we will focus on Harsanyi's preference of rule over act utilitarianism. We will not be concerned with Harsanyi's assumptions on interpersonal utility comparison, his

axiomatic justification of utilitarianism nor with the social welfare function. Hence, we will take many assumptions of his theory for granted whether we agree with them or not. It is our aim to investigate his theory and particular aspects of it, and not argue against it. It is our job as laboratory technician in the SophoLab to set up the experiment with the theory as starting point. The question of rule versus act utilitarianism is the focus point or, more precisely, which version of utilitarianism is preferable. Harsanyi is quite clear on this ([4], page 56):

"...the basic question we have to ask is this: Which version of utilitarianism will maximize social utility? Will society be better off under one or the other? This test very clearly gives the advantage to rule utilitarianism."

In Harsanyi's view the question of morality is that of maximizing social utility. This is the same for both act and rule utilitarianism. Their decision rule, however, differs. For the rule utilitarian agent the decision of other fellow rule utilitarian agents is an endogenous variable, whereas for the act utilitarian agent the decision of all others, be they utilitarian or otherwise motivated, are exogenous ([4], page 57):

"An act utilitarian moral agent assumes that the strategies of all other moral agents (including those of all other act utilitarian agents) are given and that his task is merely to choose his own strategy so as to maximize social utility when all other strategies are kept constant. In contrast, a rule utilitarian moral agent will regard not only his own strategy but also the strategies of all other rule utilitarian agents as variables to be determined during the maximization process so as to maximize social utility."

Like has been mentioned above, Harsanyi prefers rule utilitarianism over act utilitarianism. To strengthen this position, Harsanyi gives an elaborate example ([4], pages 57, 58):

"For example, consider the problem of voting when there is an important measure in the ballot but when voting involves some minor inconvenience. Suppose, there are 1,000 voters strongly favoring the measure, but it can be predicted with reasonable certainty that there will also be 800 negative votes. The measure will pass if it obtains a simple majority of all votes cast. How will the utilitarian theories handle this problem?"

First, suppose that all 1,000 voters favoring the measure are act utilitarian agents. Then each of them will take the trouble to vote only if he thinks that his own vote will be decisive in securing passage of the measure, that is, if he expects exactly 800 other people favoring the measure to vote (since in this case his own vote will be needed to provide the 801 votes required for majority). But of course, each voter will know that it is extremely unlikely that his own vote will be decisive in this sense. Therefore, most act utilitarian voters will not bother to vote, and the motion will fail (...).

In contrast, if the 1,000 voters favoring the measure are rule utilitarian agents, then all of them will vote (if mixed strategies are not allowed). This is so because the rule utilitarian decision rule will allow them a choice only between two admissible strategies: one requiring everybody to vote and the other requiring nobody to vote. As this example shows, by following the rule utilitarian decision rule people can achieve successful spontaneous co-operation in situations where this could not be done by adherence to the act utilitarian decision rule (or at least where this could not be done without explicit agreement

on coordinated action, and perhaps without an expensive organization effort).”

However, these statements immediately raise some questions. E.g., based on the assumptions made, rule utilitarian agents seem to be in a better position to organize cooperation and coordinate their strategy. However, we may wonder whether act utilitarian agents are also allowed to cooperate in one way or another. And if so, how does the coordination take place? Is there someone taking the lead and calling on others? What information flow is required for coordination to be successful? What effect would that have on their behavior? Another key issue is that of the decision rules used. Both seek to maximize utility. The main difference is that in the rule utilitarian decision rule the fellow rule utilitarian agents are endogenous, whereas in the act utilitarian decision rule all other agents’ strategies are exogenous. Regarding both decision rules the question is for both endogenous and exogenous variables: what assumptions about the values are made? And how? Do the agents form hypotheses about the behavior of others? If so, how do they reason about them? Do they apply some maximum likelihood reasoning?

We hope to have made clear that there are, at first glance, some questions (related to some ‘white spots’ in the theory) that need answering before we can say that a satisfactory explanation has been given for the claims made. We hypothesize that through the setting up of generic mechanisms to represent the problems, various options can be investigated.

4 The experiments

In this section, we prepare, model and run various experiments related to Harsanyi’s theory of utilitarianism. To do so, we apply the steps of the methodology as explained in section 2.

4.1 Step 1: Decomposition

The main components of Harsanyi’s rule and act utilitarianism as discussed in the preceding section are

- The utility functions chosen: they will be specified below in Step 2.
- The decision rule for act utilitarian agents favoring more utility over less. This concerns a decision made by each individual act utilitarian agent i yielding a strategy S_i where the agent takes into account the set of (assumed) strategies of all other agents, i.e., $S_{other} = S_1, \dots, S_{i-1}, S_{i+1}, \dots, S_z$.
- The decision rule for the rule utilitarian agents which differs in the definition of the endogenous and exogenous variables. Each rule utilitarian agent uses the same reasoning while taking into the set of (assumed) strategies of all act utilitarian agents: as a consequence of this, all rule utilitarian agents will, if subject to equal circumstances, end up with the same conclusion.
- The social welfare function $W()$ chosen, here defined as the sum of the utilities of the individual agents according to equation (1).

Harsanyi’s theory concludes with a preference of rule over act utilitarianism. Therefore, the hypothesis to be tested is the following one: rule utilitarian agents will outperform act utilitarian agents, i.e., $W_{rule} > W_{act}$ where W_{rule}, W_{act} is the social welfare function in case the agents behave like rule utilitarian agents and act utilitarian agents respectively.

4.2 Step 2: Experimental Setting

Step 2 concerns the translation of the problem into a framework and experimental setting. To do so, we take Harsanyi’s example: There are two parties, one act and one rule utilitarian party. They are engaged in a voting that each hopes to win. According to the prediction by Harsanyi the rule utilitarian agents will win the vote whenever they have the majority. And more to the point, the hypothesis is that *the act utilitarian agents will not be able to win, even if they have the majority*. Key in Harsanyi’s argument is the fact that each (act) utilitarian has two options: to go voting or do something else (that yields a positive utility). As each of act utilitarian agents has to decide for himself, the question he has to answer is: will my voting make any difference? If not, then he will do something else. As each of the act utilitarian agents will think in a similar way none will vote, and the vote is subsequently lost. The act utilitarian agent faces a situation in which there are, logically speaking, four possible outcomes. One, he will not go voting while enough of his fellows will, in which case he derives his share of the benefits from the won vote and the utility of doing X . Two, if the vote is lost he will at least have the utility from action X . Three, if he votes and the vote is lost he will derive utility from neither. Four, if the vote is won, the winning will provide utility but he has forgone the utility associated with X . To set up an executable experiment the utilities and preference functions have to be exact and quantified. We will start using the following pay-off structure defining the individual utility for the act utilitarian agents:

- do something else while enough others votes: 50
- vote while enough others votes: 40
- do something else while not enough others votes: 10
- vote while not enough others votes: 0

The rule utilitarian agents on the other hand will all follow one rule, and if that rule is to vote then they will not be in danger of unintentionally losing the voting game. So, all rule utilitarian agents will always go voting. Therefore, the pay-off structure of the rule utilitarian agents is a simple one: each rule utilitarian agent yields an individual utility of 40 in case the rule utilitarian agents win the voting, and of 0 in case they loose the voting. This discussion points to an open spot in Harsanyi’s presentation. Rule utilitarian agents are not able to exploit the majority they have. The surplus of voters cannot be used to gain additional utility doing something else, whereas the act utilitarian agents are able to do so. We finally observe here that the pay-off structures of both types of agents are parameterized and can be changed in the course of the experimentation.

4.3 Steps 3: Translation

The framework is based on the belief desire intention (BDI) model [1], implemented using the technique of computers and Java programming. It consists of agents that represent the moral agents from the Harsanyi example. As in the example they are involved in a voting. They will have the desire to maximize their utility, they holds beliefs about what other agents will do, and they will form intentions to go voting or do something else. They can reason about their beliefs and intentions in a logical way. Trying to translate this in the elements of the intermediate framework, we encounter several problems:

- In Harsanyi’s example, the agents know there are 1800 agents involved, of which 800 belong to one party and 1000 to the other party. They know that a simple majority suffices to win the vote.

As how they come to know this and other things, Harsanyi provides no explanation. Actually, Harsanyi assumes that both types of agents have certain information. E.g. with respect to the rule utilitarian agents we observe that (1) all rule utilitarian agents are supposed to be equal (that is adhere to the same decision rule), (2) they know they are the same, (3) they also know about the number of agents involved in the voting, and (4) they know what it takes to win a vote, the qualified majority.

- Perhaps this is fine for the rule utilitarian agents (since, based on these assumptions, they are able to win the voting game in many cases), but for the act utilitarian agents not so assumptions are being made. So several questions may be posed like (1) what information do act utilitarian agents know, (2) how do they get their information from? (3) how do they form their beliefs?, and (4) how do they make decisions given the choice of the other agents. To do their job, they must have some notion about their fellow agents intended behavior. Harsanyi actually does not exclude some kind of agreement or organization to arrive at some form of co-operation. Then the question is: (5) what kind of co-operation would that be? and (6) what will it cost?

From this discussion it should be clear that the argumentation as presented by Harsanyi and the example are not implementable as they are. Most crucial is the absence of an explanation how the utilitarian agents come by some knowledge about the world they live in, who their fellows are and what their intentions are. This forces to a revisiting of the steps 2 and 3.

4.4 Steps 2 till 4 (repeated): Extending and Adjusting

Let us first consider the question where the utilitarian agents get the information about the other agents from. One of the options that poses least demands on institutional organization is having the voters meet in chance groups before the vote, where they can question each other about both their type (rule or act utilitarian) and their intention (voting or not). This requires no pre-organized meetings, no knowledge of the whereabouts of fellows, no deliberate grouping (though that would certainly make sense). Moreover, it seems natural, members of a congress, members of parliament meeting each other in the corridors, the lunch room, etc. discussing the upcoming vote. In addition we observe that information collection might or might not be free. To add the possibility that information is not freely available we introduce a negative utility on enquiry. Agents are further supposed to be free to decide whether or not to acquire information. The utility (cost of information) will be first set at -2 and will be changed during the experiments.

At the start, we assume that the act utilitarian agents will hold no beliefs about other agent's intentions. They all have an innate desire to maximize utility. They go about gathering information in a group with a certain group size. Typically, the groups for the information exchange are much smaller than the voting group (which consists of all potential voters). The information exchange groups are assembled at random. Agents can be added to groups as a member. Agents decide whether they want to participate in the information exchange and voting and sign up for a group. In the experiment, all agents will participate in the information exchange where they ask their fellows about their intentions. Now the question is where does this first intention come from. If each agent asks the other agents about their intention, needing this information to form his own intention, we are stuck. We therefore assume some propensity to go voting. This

propensity will be modeled as the chance to go voting (i.e., the complementary chance will express the probability to do something else). Given the propensity probability, the initial intention to go voting (or not) is assigned randomly to the (act) utilitarian agents. Next, each agent i contacts the members of the information exchange group they belong to, in order to form his (her) personal belief about the expected number $E_i(V)$ of voters of each party (excluding agent's own vote). In its most basic form this is done by extrapolating the findings in the small groups to the whole population.

Given their beliefs, the next question is how the act utilitarian agents will form their final intention (that is transformed into action). To start with, there are three different decision rules that come to mind where α represents the majority vote, i.e., for our voting game with 1800 voters, $\alpha = 901$.

1. Harsanyi's version: if agent i thinks he will make a difference go voting, otherwise do something else or, more formally:

$$\begin{aligned} &\text{Go voting if } \alpha - x < E_i(V) < \alpha + y \text{ where} \\ &\quad x = 2 \text{ and } y = 0, \\ &\text{otherwise do something else.} \end{aligned} \quad (3)$$

2. Generalized version: if agent i thinks his party will win or lose anyway do not go voting but use the opportunity to gain some additional utility by doing something else or, more formally:

$$\begin{aligned} &\text{Go voting if } \alpha - x < E_i(V) < \alpha + y \text{ where} \\ &\quad \alpha - x \in [0, 1800] \text{ and } \alpha + y \in [0, 1800], \\ &\text{otherwise do something else.} \end{aligned} \quad (4)$$

3. An even more extended version which is expressed as follows:

$$\begin{aligned} &\text{Stick with intention if } \alpha - x < E_i(V) < \alpha + y, \\ &\text{do something else if } E_i(V) \geq \alpha + y, \\ &\text{and go voting if } E_i(V) \leq \alpha - x, \\ &\quad \text{where } \alpha - x \in [0, 1800] \text{ and } \alpha + y \in [0, 1800]. \end{aligned} \quad (5)$$

Rule 1 (defined by (3)) is Harsanyi's version of the act utilitarian decision rule. If, and only if, the agent's vote is decisive, he will go voting. Rule 2 (defined by (4)) is a generalized version of rule 1. The only difference being that the margins within which the agent will conceive his vote as decisive is extended. He does not have to be voter 901 in order to go voting but might be say voter 904. A justification for this extension is uncertainty. Under circumstances it might be hard to get a correct impression of the exact number of voters. One might get the impression wrong by some margin. This margin then has to be taken into account. Rule 3 defined by (5) is different. It is introduced as an alternative in the experiment to see whether other outcomes are possible under different assumptions. It takes the current intention of the agent into account as well as the expected outcome of the voting based on the expected number of voters of each party. If the expectation is that the vote will probably be won, the current intentions by all agents is perfectly suited and should not be altered, including its own one. Again, there are some margins for uncertainty. If the expectation is that the vote will be won by a sufficiently large margin, the agent will decide to do something else. If the expectation is that the vote will be lost this has to be remedied by going to vote. This is probably the hardest element in the decision rule to justify from a rational point of view. The chance that the vote will be decisive is small indeed, the utility of doing something else is guaranteed and should be preferred. Otherwise, if the alternative utility is small, the agent may take the chance and decide to go voting.

4.5 Step 5 and 6: Implementation

Trying to keep the focus on the main arguments has made us decide to skip all kinds of implementation details. Here we only mention that, as basic elements of the SophoLab framework, so-called agents, groups, games, and a referee are used to implement the above-sketches ideas. For more details, we refer to [6], chapter 5.

4.6 Step 7: Running the experiments

Running the experiments consists of executing the computer program with different sets of the parameters. The model embedded in the software has a set of input and output variables. The input variables include the decision rules, the number of runs, the tolerance margins, the number and types of agents, the propensity (to vote or not to vote), the group size (of inquiry groups), the sequence of actions, and the pay-off matrices. The output variables are the average score of all types of agents, as well as their max, min scores.

Table 1. The results of 6 voting runs with 48 act utilitarian and 32 rule utilitarian agents: the first six rows show the input parameter values, the last two rows the resulting output in terms of average score (av sc), maximum score (max sc), and minimum score (min sc).

decision rule	3	number of runs	6
margins	3.5, 3.5	number of agents	80
agent types	act, rule utilitarian	number per type	48, 32
propensity	0.8	inquiry group size	40
sequence	2x inquiry + voting	pay-off inquiry	1
		pay-off voting	50, 40, 10, 0
av sc (act)	243	max, min sc (act)	288, 228
av sc (rule)	0	max, min sc (act)	0, 0

4.6.1 Some first results

Table 1 represents the results of a run in which 80 agents, of which 48 act utilitarian agents and 32 rule utilitarian agents are involved in a voting. The act utilitarian agents follow decision rule 3 (as defined by (5)) with (tolerance) margins of 3.5. This means that if they expect between 37,5 (38) and 44,5 (44) of their fellows to go voting, they will stick to their original intention. The propensity (probability of going to vote, initially) equals 0.8. Gathering information costs 1 utility for act utilitarian agents (not of relevance for rule utilitarian agents) and is done in groups of 40 agents (defined by the inquiry group size), winning the vote brings 40 utilities per agent, etc. The outcomes of this experiment are as follows: on average the act utilitarian have a score of 243 which is slightly more than the rule utilitarian could have achieved. The maximum average utility for the rule utilitarian agents is 240 (6 times 40 for winning the vote when they are the majority party). The act utilitarian that was best had a total utility of 288 while the worst off scored 228.

Following decision rule 2 with wider tolerance margins shows similar results in case the total number of agents is 20, with 12 act utilitarian agents and 8 rule utilitarian agents. The group size in which information is gathered was smaller (namely 5 which equals 25% of the total group size). On average the utility gathered by the act utilitarian agents appeared to be slightly less than what rule utilitarian agents would have achieved. Most importantly, we again observed that the rule utilitarian agents could never win the voting: for further details we refer to [6].

4.6.2 General findings

Many runs were executed in which some variables were kept constant while one or two other variables were varied to investigate its success. By running many such tests a picture arises that we will now describe. The experiments show that, independent of the configuration, decision rule 1 (as described by inequality (3)) is disastrous for the act utilitarian agents. They never win a voting, not even when they form the majority, as predicted by Harsanyi. When the decision rule is relaxed in order to include uncertainty, the act utilitarian agents fare better. In some runs they are able to win the voting. Important seems to be the tolerance in the decision rule, that is the extent of uncertainty allowed for. Decision rule 3 is even more successful. From a fairly small tolerance onwards the act utilitarian agents are able to win the vote when they have the majority. All decision rules allow the act utilitarian agents to exploit the majority surplus. Part of the population does not vote while the vote is still won. In cases where the vote is lost, still some utility is gained by some (rule 2 and 3) or all act utilitarian agents (rule 1).

The tolerance margin can vary from zero to half the size of the population. With a tolerance of zero the decision rule is the one proposed by Harsanyi. With a tolerance of half the population size we have effectively a rule that says ‘vote always’, this is, of course, the rule utilitarian strategy. As Harsanyi predicted with a tolerance of zero, act utilitarian agents are not able to win a vote. What did surprise was that after an increase to about 3.5, act utilitarian agents are almost always winning the vote when they have the majority. Another important element is the cost of information. From the previous aspect of tolerance we learned that some tolerance in the decision making helps. This is, of course, only the case if there is some information about what to expect. Thus information exchange is vital. Information is valuable only if it helps increase the chances of a won vote, which again is in part dependent on the tolerance. As the cost of information increases act utilitarian agents still win their votes, but at an increasing cost. When cost is high rule utilitarian agents do markedly better because they have less need for information. This relationship is directly proportional.

4.7 Steps 8 and 9: Translating back to the theory

The decision rule as described by Harsanyi works out badly for the act utilitarian agents. The generalized version (decision rule 2) works already better while the adapted version (decision rule 3) proves even more beneficial. We argued above that the adaptation of the rule does not violate the act utilitarian character, but does take into account uncertainty (which is left out of Harsanyi’s account). So with a slight relaxation of the decision rule act, utilitarian agents win the vote, contrary to Harsanyi’s prediction. And under certain conditions - larger tolerance margins - we have seen that act utilitarian agents perform better than rule utilitarianism could have done. This follows from their ability to exploit the surplus of votes.

The size of the informal group that exchange information influences the performance significantly. The relationship is not linear. Small (12,5% of total population) and large (50% of total population) groups perform clearly better than medium sized (25% of total population) groups. As the population size grows act utilitarian agents improve their performance. For rule utilitarian agents the minimum and maximum scores are always the same. For the act utilitarian agents the minimum and maximum score vary markedly. The individual differences are explained by the random influences that are built in through both inclination and grouping. The decision rule

appears to be fairly stable to variations in propensity to vote among the act utilitarian agents.

There are stable decision rules that allow act utilitarian agents to function successfully with a minimal requirement of information. The situations in which act utilitarian agents outperform rule utilitarian agents are by no means artificial. The success of act utilitarian agents depends to an important extent on the availability and costs of information, and on the decision rule. Contrary to Harsanyi's claim act utilitarian agents can successfully and spontaneously coordinate their actions. This requires a somewhat different decision rule.

5 Discussion and Conclusions

We have examined Harsanyi's arguments in favor of rule utilitarianism over act utilitarianism. His arguments stand within a heated (old) discussion on which version of utilitarianism is to be preferred. His claim that "...the basic question we need to ask is this: Which version of utilitarianism will maximize social utility? Will society be better off under one or the other? This test very clearly gives the advantage to rule utilitarianism" was tested in an experimental setting. And, not only does Harsanyi make a particular claim, he also provides the criterion by which the claim is to be judged: maximization of social utility. Since social utility is defined as the basis of the morally good in utilitarianism, our (experimental) approach of calculating social utility scores for different utilitarian rules is in the heart of the philosophical discussion on what is the preferred version.

The experiments executed show that act utilitarian agents need not fare worse than rule utilitarian agents in certain circumstances. This is especially remarkable if one takes into account that they can achieve their results by epistemically less demanding assumptions. They are also able to exploit the surplus of votes when they have the majority to gain some additional utility. This compensates for their occasional loss of the vote due to imperfect (wrong) expectations about the number of fellow act utilitarian that will show up. Core to this ability to perform fairly well is a small relaxation of the decision rule as presented by Harsanyi. It consists of allowing some degree of uncertainty into the decision rule.

The experiments we ran are limited in scope and are open to several objections. Actually, several other assumptions may be chosen with respect to both act utilitarian agents and rule utilitarian agents. We are aware of the fact that the corresponding experiments may yield still other outcomes. We have planned to perform such additional philosophical experiments in the near future. So, at this moment, none of the conclusions and observations we made in this paper are conclusive. But at least we hope to have shown - and that is the main message based on the research performed so far - that setting up experiments is a useful way to gain new and deeper insights in existing argumentations used in old (and new) philosophical debates.

References

- [1] M.E. Bratman, *Intention, Plans and Practical Reasoning*, Harvard University Press, Cambridge, 1981.
- [2] P. Danielson, *Artificial Morality*, Routledge, London, 1992.
- [3] *Modeling Rationality, Morality and Evolution*, ed., P. Danielson, Oxford University Press, New York, 1998.
- [4] J.C. Harsanyi, 'Morality and the Theory of Rational Behaviour', in *Utilitarianism and Beyond*, ed., Sen et al., (1982).
- [5] John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ, 3rd edn., 1953.
- [6] Vincent Wiegel, *SophoLab, Experimental Computational Philosophy*, Ph.D. dissertation, Faculty of Technology, Policy and Management, Delft University of Technology, 2007.