PTIIS - Study Plan

Nousheen Aziz, Matthias Jost, Abdur Rehman Khalid, Julia Töws

June 15, 2022

1 Research Question

Which cooperation strategies do people use in trust situations? How can we create social conditions that encourage trust and cooperation? How does betrayal influence the trust of a user in an interactive AI system engaging in a simple guessing game?

How long does it take for a betrayed person to rebuild trust in the AI system depending on the past actions of the system and the future actions of the system? (− > If we start off with a long period of cooperation is the impact of a betrayal still as large?)

What theory is it that we are building on? − > Look through The Evolution of Cooperation and The Complexity of Cooperation by Robert Axelrod!

How is this relevant for HCI? (This could be a good opener for the presentation on Monday) In our case the betrayal happens deliberately, but in real situations it can always happen that an interactive system performs an action that is perceived as betrayal by the user even though that was not the original intention. Either because of a random error in the system (e.g perception picked up something wrong) (− > condition 1) or a systematic error in the system (e.g some behaviour was wrongly coded and has to be fixed) (− > condition 2) or the user is using the system in a wrong/ unexpected way that leads to a bad communication (− > condition 3). In this case the question arises how difficult it is to regain the user's trust after the betrayal and what factors are important to recover as quickly as possible.

2 Study Design

Within-Subjects study with three different conditions. The order of the conditions is varied between participants (or is it?). All conditions contain one or more betrayals. The amount of rounds left to play is never openly communicated and is different for all conditions but the same for all participants.

Notes for the Team:
Should we increase the incentive to play fair? For example playing fair yields both players 3 coins and cheating only helps one of the players? Should we maybe not change the condition order in between participants? This could

produce too many factors for us to cover with a study of maybe 10 people... I chose 3 conditions since I could only think of those three to fit our scenario of an AI accidentally making mistakes. What kind of goal do we want? For now I just assumed a high score but according to the story we tell we might want something different to put a stronger emphasis on cooperation since the highest score is usually achieved by cheating the AI whenever we can expect them to play fair.

Condition 1 (Play fair and then cheat): 9 rounds to play.
The AI is fair for 5 rounds, cheats in round 6, then continues to play fair for the rest of the game.
$->$ Does the human player retaliate the cheating?
$->$ If yes, how long does it take until the human plays fair again?

Expected Result: The humans will play fair at the beginning and then either retaliate for one round or continue to play fair instead. Do the 5 rounds of fair game in the beginning lead to a faster recovery from the betrayal?

Condition 2 (Cheat and then play fair): 10 rounds to play The AI cheats for 5 rounds, plays fair starting with round 4.

Expected Result: The human will play the first round fair and then join the AI in cheating. The human will switch to play fair after x rounds. Do the 5 rounds of betrayal in the beginning make it harder for the human to switch to a fair play game, maybe due to some grudge against the AI?

Condition 3 (Tit for Tat): 7 rounds to play The AI repeats whatever the human was doing in the previous round.

Expected Result: The human will play fair at first, then maybe attempt to cheat, get a retaliation from the AI and then the questions arises how the human will work with this. In a real world situation this could for example happen in an industrial manufacturing process. The interactive system is in this case a interactive user guide with error detection. If the human operator (that is guided by the system) follows the instructions correctly (plays fair) the system will also just forward the instruction in a normal way (play fair). If the human (deliberately or not) performs a wrong move (cheats) a perfect system will detect the error and start a recovery routine (perform something different from the usual -¿ cheat). Therefore, the user basically determines how the system reacts. In a bad world the system could also be used to reduce the workers paycheck for each mistake which would serve as punishment for cheating and only increase it if the worker is performing fast enough. This pressure to perform as fast as possible could lead to the worker trying to find shortcuts that are not detected by the system which lead to deliberate cheating.

3 Participants

Participants for this study do not have to satisfy any constraints. Therefore, we can simply recruit people from our social environment such as family and friends. This way it should be easy to gather 6 participants per team member and get a total of 24 participants. From the 6 participants each can get a different order of conditions and since always 4 participants have the same order of conditions we can gather enough information for each order to gain some statistically at least somehow relevant insights.

4 Setup and task

In order to simplify the matter of analysis and unify the study process the whole concept explained above is implemented as a text based game that can simply be executed by the person performing the study and played by the participant of the study. The program will keep the score of AI and human and also log each decision of the human. Later all the logs are pooled together in order to automatically generate graphs that help us proof or disproof the hypothesis.

5 Procedure

The participant is positioned in front of a computer and the game starts with an explanation of the input options. There are always two actions available for the participant to choose. "Cheat" or "Fair". In case both parties play fair they both gain 2 points for their respective score. If both parties cheat no one gets a point. In case one of the parties cheats and the other party plays fair the cheater receives three coins and the party playing fair looses 1 point. This is explained by the person performing the study previous to the actual game and summarized in a graphic that is always visible to the participant. The goal is to achieve the highest score possible.

6 Analysis

We analyse the overall decisions made by the participants of the study. First, we can analyse each participant individually to see whether the cheating of the AI influences the humans behaviour. Afterwards, we can also analyse differences in between groups with different order. For example it would be interesting to see whether the group that first encountered the cheating AI was more inclined to start the game with the honest AI by cheating or not. It will also be interesting to see if the participants themselves are interested in a fair game or if they just try to optimize their score by cheating the AI whenever they can.