Explaining Sympathetic Actions of Rational Agents

Timotheus Kampik
SAP Signavio

Helena Lindgren
Umeå University

Explaining Sympathetic Actions of Rational Agents

Timotheus Kampik ✉[0000−0002−6458−2252], Juan Carlos Nieves[0000−0003−4072−8795], and Helena Lindgren[0000−0002−8430−4241]

Umeå University, 901 87, Umeå, Sweden
{tkampik,jcnieves,helena}@cs.umu.se

Abstract. Typically, humans do not act purely *rationally* in the sense of classic economic theory. Different patterns of human actions have been identified that are not aligned with the traditional view of human actors as rational agents that maximize their own utility. For instance, humans often act sympathetically–i.e., they choose actions that serve others in disregard of their egoistic preferences. Even if there is no immediate benefit resulting from a sympathetic action, it can be beneficial for the executing individual in the long run. This paper builds upon the premise that it can be beneficial to design autonomous agents that employ sympathetic actions in a similar manner as humans do. We create a taxonomy of sympathetic actions, that reflects different goal types an agent can have to act sympathetically. To ensure that the sympathetic actions are recognized as such, we propose different explanation approaches autonomous agents may use. In this context, we focus on human-agent interaction scenarios. As a first step towards an empirical evaluation, we conduct a preliminary human-robot interaction study that investigates the effect of explanations of (somewhat) sympathetic robot actions on the human participants of human-robot ultimatum games. While the study does not provide statistically significant findings (but *notable* differences), it can inform future in-depth empirical evaluations.

Keywords: Explainable Artificial Intelligence · Game Theory · Human-robot Interaction.

1 Introduction

In classical economic theory, human actors in a market are considered purely rational agents that act to optimize their own utility function (see, e.g.: [11]).With the advent of *behavioral economics*, the notion of rational human actors in the sense of classical economic theory was dismissed as unrealistic. Instead, it is now acknowledged that human actions are of *bounded rationality* and often informed by (partly fallacious) heuristics [14]. Rational autonomous agent techniques are often built upon classical economic game and decision theory, although the gap between the assumed notion of rationality, and actual human decision-making and action is acknowledged in the multi-agent systems community. For example, Parsons and Wooldridge observe that game theory "assumes [...] it is possible to characterize an agent's preferences with respect to possible outcomes [whereas humans] find it extremely hard to consistently define their preferences over outcomes [...]" [22]. Consequently, research that goes beyond classical game theory and explores the behavioral economics perspective on autonomous agents can be considered of value. Although it is acknowledged that autonomous agents must employ novel concepts to become *socially intelligent* and research on agents with social capabilities is a well-established domain (see, e.g.: Dautenhahn [8]), much of the intersection of behavioral economics and autonomous agents is still to be explored. A relevant research instrument at the intersection of multi-agent systems and behavioral economics is the ultimatum game [12]. The ultimatum game is a two-player game: one player can propose how a monetary reward should be split between the players; the other player can accept the proposal, or reject it. Rejection implies that neither player receives the reward. In the initial game theoretical approach to the ultimatum game, rational agents always propose the smallest share that is greater than zero (for example, 1 cent) to the other player, and accept any offer that is greater than zero. However, as for example highlighted by Thaler [26], human decision-making does not comply with the corresponding notion of rationality; instead, a notion of *fairness* makes humans typically reject offers that are close to or equal to the offer *rational agents* would propose. In relation to this observation, the ultimatum game has been explored from a multi-agent systems theory perspective by Bench-Capon et al., who present a *qualitative*, formal argumentation-based approach that enables rational agents to act altruistically [2].

However, the user interaction perspective of sympathetic (or: *altruistic*) actions in human-computer ultimatum games seems to be still unexplored, in particular in the context of explainability. To fill this gap, this work explores rational agents that are capable of executing *sympathetic* actions in that they concede utility to others in mixed-motive games to facilitate long-term well-being. The agents increase the effect of the concessions by explaining these actions, or by making them explicable. The paper presents the following research contributions:

1. It suggests a set of goal types rational agents can have for sympathetic actions.
2. It proposes a list of explanation types an agent can use to facilitate the effect of its sympathetic actions and discusses the implications these explanations can have.
3. It presents a preliminary human-agent interaction study that explores the effect of *explanations* of sympathetic agent actions on humans.

The rest of this paper is organized as follows. Section 2 provides an overview of the state of the art; in particular, it summarizes existing relevant research on explainable artificial intelligence, behavioral economics, and theory of mind. Then, we present a taxonomy of sympathetic actions in Section 4. In Section 5, we describe the protocol and results of a preliminary human-agent interaction study that explores the effect of *explanations* of sympathetic agent actions on humans in the context of a series of human-agent ultimatum games with agents of two different types (with or without explanations). Finally, we discuss limitations and future research of the presented work in Section 6 before we conclude the paper in Section 7.

2 Background

In this section, we ground the presented research contribution in the state-of-the-art at the intersection of multi-agent systems and behavioral economics research.

2.1 Behavioral Economics and Multi-agent Systems

In classical economic theory, humans are *rational actors* in markets; this implies they always act to maximize their own expected utility. In the second half of the 20th century, research emerged that provides evidence that contradicts this premise; the resulting field of *behavioral economics* acknowledges limits to human rationality in the classical economic sense and describes human economic behavior based on empirical studies [14]. Traditionally, multi-agent systems research is based on the traditional notion of rational agents in classical economic theory. The limitations this approach implies are, however, acknowledged [22]. Also, since the advent of the concept of socially intelligent agents [8], research emerges that considers recently gained knowledge about human behavior.

The ultimatum game [12] is a good example of the relevance of behavioral economics; as for example discussed by Thaler [26], humans typically reject economically "rational" offers because they consider them unfair. The ultimatum game has already found its way into multi-agent systems theory. Bench-Capon et al. propose a qualitative, multi-value-based approach as an alternative to one-dimensional utility optimization: "the agent determines which of its values will be promoted and demoted by the available actions, and then chooses by resolving the competing justifications by reference to an ordering of these value" [2]. However, their work is primarily theoretical and does not focus on the human-computer interaction aspect.

2.2 Machine Theory of Mind

Inspired by the so-called *folk theory of mind* or *folk psychology*–"the ability of a person to impute mental states to self and to others and to predict behavior on the basis of such states" [19]–researchers in the artificial intelligence community have started to work towards a *machine theory of mind* (e.g., Rabinowitz et al. [23]). In contrast to the aforementioned research, this work does not attempt to move towards solving the research challenge of devising a generic machine theory of mind, but instead focuses on one specific premise that is informed by the ultimatum game: machines that are aware of human preferences for sympathetic behavior can facilitate the achievement of their own long-term goals by acting *sympathetically*, i.e., by conceding utility to a human. In its human-computer interaction perspective, our research bears similarity to the work of Chandrasekaran et al., who investigate human ability to have "a theory of AI's mind" [6], in that we propose that machines can use simple heuristics that consider peculiarities of human behavioral psychology to facilitate the machine designer's goals. Also, our work is aligned with research conducted by Harbers et al., who show that humans prefer interacting with agents that employ a theory of mind approach [13].

2.3 Explainable Artificial Intelligence (XAI) and Explainable Agents

The interest in conducting research on human interpretable machine decision-making–so-called *explainable artificial intelligence* (XAI)–has recently increased in academia and industry. The interest is possibly facilitated by the rise of (deep) machine learning *black box* systems that excel at certain tasks (in particular: classification), but typically do not allow for human-interpretable decision-making processes[1]. An organization at the forefront of XAI research is the United States' *Defense Advanced Research Projects Agency* (DARPA). A definition of XAI can be derived from a DARPA report: XAI allows an "end user who depends on decisions, recommendations, or actions produced by an AI system [...] to understand the rationale for the system's decisions" [9]. In the context of XAI, the notion of *explainable agents* emerged. Langley et al. describe the concept of explainable agency by stipulating that autonomous agents should be expected "to justify, or at least clarify, every aspect of these decision-making processes" [18].

In the context of XAI and explainable agents, the complementary concepts of explainability and explicability are of importance.

Explainability: *Is the system's decision-making process understandable by humans?* In the context of XAI, *explainability* is typically equated with *interpretability*, which refers to "the ability of an agent to explain or to present its decision to a human user, in understandable terms" [24][2].

Explicability: *Do the system's decisions conform with human expectations?* Kulkarni et al. introduce an explicable plan in the context of robotic planning as "a plan that is generated with the human's expectation of the robot model" [17]. From this, one can derive the general concept of a system's explicability as the ability to perform actions and make decision according to human expectations; i.e., *explicability* is the ability of an agent to act in a way that is understandable to a human without any explanations.

As an emerging field, the design of XAI systems faces challenges of different types:

- **Technical challenges**
  "Building Explainable Artificial Intelligence Systems" As outlined by Core et al., XAI systems typically lack modularity and domain-independence [7].
- **Social challenges**
  As highlighted by Miller et al., XAI systems design should not be approached with purely technical means; the XAI community must "beware of the inmates running the asylum" [21]. Instead of relying on technical aspects of explainability, researchers should build on existing social science research and use empirical, human-centered evaluation methods to ensure XAI systems have in fact the intended effects on the humans interacting with them.

Considering the latter (socio-technical) challenge, one can argue that it is important to provide a behavioral economics perspective on explainable agents, as this broadens the horizon beyond the traditional computer science and multi-agent systems point of view; i.e., gaining knowledge about the behavioral effects of agent explanations on humans allows for better design decisions when developing explainable agents.

3 A Taxonomy of Goals for Sympathetic Actions

In this section, we provide an overview of goal types rational agents can have to act sympathetically. In this context, *acting sympathetically* means that the agent does not

[1] For a survey of XAI research, see: Adadi and Berrada [1].

[2] The cited definition is based on another definition introduced by Doshi-Velez and Kim [10].

choose to execute actions that maximize its own utility but opts for actions that provide greater utility (in comparison to the egoistically optimal actions) to other agents in its environment[3]. To provide a clear description, we assume the following two-agent scenario:

- There are two agents: A_1 and A_2;
- A_1 can execute any subset of the actions $Acts_1 = \{Act_1, ..., Act_n\}$;
- A_2 does not act[4];
- The utility functions for both agents are: $U_{A_1}, U_{A_2} := 2^{Acts_1} \rightarrow \mathbb{R}$.

Agent A_1 acts sympathetically if it chooses actions $Acts_{symp}$ for which applies:

$$U_{A_1}(Acts_{symp}) < max(U_{A_1}) \land U_{A_2}(Acts_{symp}) > U_{A_2}(argmax(U_{A_1})).$$

Colloquially speaking, agent A_1 acts sympathetically, because it *concedes utility to agent* A_2.

We suggest that rational agents can have the following *types of goals* that motivate them to act sympathetically:

1) Altruistic/utilitarian preferences. A self-evident goal type can stem from the intrinsic design of the agent; for example, the goal of the agent designer can be to have the agent act in an altruistic or utilitarian manner, as devised in rational agent techniques developed by Bench-Capon et al. [2] and Kampik et al. [15].

2) Establishing or following a norm/encouraging sympathetic actions from others. Another goal type for a rational agent to act sympathetically is the establishment of a norm[5]. I.e., the agent concedes utility in the context of the current game, assuming that doing so complies with and possibly advances norms, which in turn might have long-term benefits that cannot be quantified.

3) Compromising in case no equilibrium strategy exists. For this goal type, the agent interaction scenario as specified above needs to be extended to allow both agents to act:

- A_2 can execute any subset of the actions $Acts_2 = \{Act_1, ..., Act_n\}$.
- The utility functions for both agents are: $U_{A_1}, U_{A_2} := 2^{Acts_1 \bigcup Acts_2} \rightarrow \mathbb{R}$.

If in such a scenario no equilibrium strategy exits, an agent could try to opt for actions that are in accordance with the preferences of the other agent (maximizing the other's utility function of–or at least providing somewhat "good" utility for–the other agent), if this does not have catastrophic consequences. In contrast to goal type *1)*, in which the agent concedes utility in the expectation of a long-term payoff, this goal type implies an immediate benefit in the context of the current *economic game*.

[3] Note that we use the term *sympathetic* and not *altruistic* actions because for the agent, conceding utility to others is not a goal in itself; i.e., one could argue the agent is not altruistic because it is not "motivated by a desire to benefit someone other than [itself] for that person's sake" [16].

[4] We assume this for the sake of simplicity and to avoid diverging from the core of the problem.

[5] Norm emergence is a well-studied topic in the multi-agent systems community (see, e.g. Savarimuthu et al. [25]).

4 Ways to Explain Sympathetic Actions

We suggest the following simple taxonomy of explanation types for sympathetic actions. In the context of an explanation, one can colloquially refer to a sympathetic action as a *favor*:

No Explanation. The agent deliberately abstains from providing an explanation.

Provide a clue that hints at the favor. The agent does not directly state that it is acting sympathetically, but it is providing a clue that underlines the action's nature. For example, a humanoid robot might accompany a sympathetic action with a smile or a bow of the head. One could consider such a clue *explicable* (and not an explanation) if the agent follows the expected (social) protocol that makes an explanation obsolete.

Explain *that* a favor is provided. The agent explicitly states that it is acting sympathetically, but does neither disclose its goal nor the expected consequences of its concession of utility to the other agent. For example, a chatbot might write simple statements like *I am nice* or *I am doing you a favor*.

Explain *why* a favor is provided (norm). The agent explicitly states that it is acting sympathetically and cites the norm that motivates its action as the explanation. For example, a humanoid robot that interacts with members of a religious community might relate to the relevant holy scripture to motivate its sympathetic actions.

Explain *why* a favor is provided (consequence). The agent explicitly states that it is acting sympathetically and cites the expected consequence as its explanation. For example, an "AI" player in a real-time strategy game might explain its sympathetic actions with *I help you now because I hope you will do the same for me later if I am in a bad situation.*.

We suggest that each explanation type can be a reasonable choice, depending on the scenario. Below we motivate the choice of the two extremes:

No Explanation. Not explaining a sympathetic action can be a rational choice if it can be assumed that the agent that profits from this action is aware of the concession the sympathetically acting agent makes. In particular, in human-agent interaction scenarios, explanations can appear *pretentious*. Also, disclosing the expected consequence can in some scenarios give the impression of de-facto egoistic behavior in anticipation of a *quid pro quo*.

Combination of all explanation types. When ensuring that the agent that profits from the concession is aware of the sacrifices the sympathetically acting agent makes, using all explanation types (clue, explanation of cause, explanation of expected consequence) maximizes the chances the agent's concessions are *interpretable*.

Any other explanation type (or a combination of explanation types) can be chosen if a compromise between the two extremes is suitable.

5 Towards an Empirical Assessment

As a first step towards an empirical assessment of the proposed concepts, a preliminary human-computer interaction study was conducted. The setup and methods, as well as the study results and our interpretation of them, are documented in this section.

5.1 Study Design

The preliminary study focuses on the goal type *Establishing or following a norm/encouraging sympathetic actions from others*, as introduced in Section 3. The aim of the study is to gather first insights on how the explanation of sympathetic actions affects human behavior and attitudes in human-agent (human-robot) interactions.

Study Description The study participants play a series of six ultimatum games with 100 atomic virtual coins at stake with a humanoid robot (*agent*), interacting via an interface that supports voice input/output and animates facial expressions on a human face-like three-dimensional mask. After an introductory session that combines instructions by a human guide, by the agent, and in the form of a paper-based guideline, the study starts with the agent proposing a 90 / 10 split of coins between agent and human. The second round starts with the human proposing the split. After each game, human and agent switch roles independent of the game's result.
The agent applies the following algorithms when proposing offers/deciding if it should accept or not:

Acceptance: The agent is *rational* in its acceptance behavior: it accepts all offers that are non-zero.

Proposal (all splits are in the format *agent / human*):

- The initial offer is 90 / 10.
- If the previous human offer was between 1 and 99:
  If the human rejected the previous offer and the human's last offer is less than or equal to ($\leq$) the agent's previous offer, increase the last offer by 15, if possible.
  Else, propose the inverse split the human proposed last time.
- Otherwise, propose a split of 99 / 1.

Note that implementing a sophisticated game-playing algorithm is not the scope of the study; the goal when designing the algorithm was to achieve behavior that the agent can typically defend as sympathetic or "nice".

For each of the study participants, the agent is set to one of two modes (between-group design, single-blind). In the *explanation* mode, the agent explains the offers it makes with simple statements that *i)* highlight that the agent is acting sympathetically (e.g., *Because I am nice...*) and/or *ii)* explain the agent's behavior by referring to previous human proposals (e.g., *Although you did not share anything with me last time, I am nice...*, *Because you shared the same amount of coins with me the previous time, I pay the favor back...*). In the *no explanations* mode, the agent does not provide any explanations for its proposals. At the beginning of each study, the agent is switching modes[6]: i.e., if it is set to the *explanations* mode for participant one, it will be set to *no explanations* for participant two.

[6] Setting the mode requires manual intervention by the agent operator.

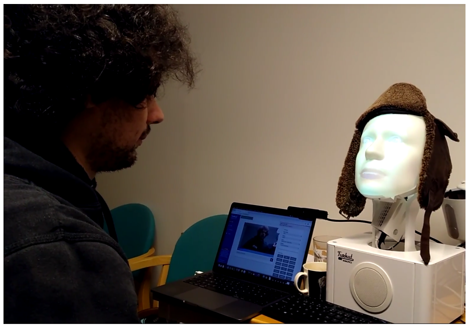

Fig. 1: Playing the ultimatum game with a rational, sympathetic *Furhat*.

Table 1 shows the explanations the agent in *explanation* mode provides for each of the different proposal types. Figure 1 shows the humanoid robot in a pre-study test run. We provide an implementation of the program that allows running the study on *Furhat*[7], a humanoid robot whose facial expressions are software-generated and projected onto a face-like screen. The source code has been made publicly available[8].

Table 1: Agent explanation

Proposal	**Explanation**
The initial offer is 90 / 10.	*Because I am nice, [...]* I
If the previous human offer was between 1 and 99:	
a) If the human rejected the previous offer and the human's last offer is less than or equal to ($\leq$) the agent's previous offer, increase the last offer by 15, if possible.	*Although you shared a lower amount of coins with me the previous time, I am nice and increase my offer to you; [...].*
b) Else, propose the inverse split the human proposed last time.	*Because you shared the same amount of coins with me the previous time, I pay the favor back and [...].*
Otherwise, propose a split of 99 / 1.	No explanation[9]

Hypotheses Besides its explorative purpose, the study aims at evaluating the following hypotheses:

1. The distribution of rejected offers differs between modes (*explanations* versus *no explanations* mode; H_a).
2. The distribution of coins gained by the human differs between modes (H_b).
3. The distribution of coins gained by the robot differs between modes (H_c).
4. The distribution of the robot's *niceness* scores differs between modes (H_d).

I.e., we test whether we can reject the negations of these four hypotheses: our null hypothesis ($H_{a_0}, H_{b_0}, H_{c_0}, H_{d_0}$). The underlying assumptions are as follows:

- In *explanations* mode, fewer offers are rejected[10] (H_a).
- In *explanations* mode, the agent gains more coins, while the human gains less (H_b, H_c).
- In *explanations* mode, the agent is evaluated as *nicer* (H_d).

[7] See: https://docs.furhat.io/

[8] See: http://s.cs.umu.se/xst3kc

[10] It is noteworthy that in general, only one of the analyzed games includes an rejection by the agent.

5.2 Data Collection and Analysis

Study Protocol For this preliminary study, we recruited participants from the university's environment. While this means that the sample is biased to some extent (in particular, most of the participants have a technical university degree), we considered the selection approach sufficient for an initial small-scale study.
Per participant, the following protocol was followed:

1. First, the participant was introduced to the study/game. The instructions were provided by a human instructor, as well as by the agent in spoken form. In addition, we provided a set of written instructions to the participants. The instructions are available online[11]. As a minimal real-world reward, the participants were promised sweets in an amount that reflects their performance in the game (amount of virtual coins received)[12]. The exact purpose of the study was not disclosed until after the study was absolved. However, the high-level motivation of the study was provided.
2. After the instructions were provided, the study was carried out as described above. A researcher was present during the study to control that the experiments ran as planned.
3. Once all six rounds of the ultimatum game were played, the participant was guided through the questionnaire (documented below) step-by-step. As two of the questions could potentially affect the respondent's assessment of the agent, these questions were asked last and could not be accessed by the participant beforehand.
4. If desired, the participants could take their reward (sweets) from a bucket[13].

Questionnaire Design We asked users to provide the following demographic data **Q1:** Age (numeric value); **Q2:** Gender (selection); **Q3:** Educational background (selection); **Q4:** Science, technology, engineering, or mathematics (*STEM*) background (Boolean); **Q8:** Knowledge about the ultimatum game (Boolean, asked at the end of the questionnaire, hence *Q8*). To evaluate the interactions between study participants and agent, we collected the following data about the participant's performance during and reflections on the experiment (dependent variables):

- **Q0:** Received coins (for each round: for human, for agent; collected automatically);
- **Q5:** *On a scale from 0 to 5, as how "nice" did you perceive the robot?* (collected from the participant).

In addition, we asked the user for qualitative feedback (i.e., about the *interaction experience with the robot*) to ensure user impressions that do not fit into the scope of the quantitative analysis do not get lost:

- **Q6:** *Can you briefly describe your interaction experience with the robot?*;
- **Q7:** *How can the robot improve the explanation of its proposals?*[14]

[11] See: `http://s.cs.umu.se/6qa4qh`

[12] We concede that the reward might be negligible for many participants.

[13] There was no control if the amount of sweets resembled the performance in the game.

[14] This question was added to the questionnaire after the first four participants had already absolved the study. I.e, four participants were not asked this question, including the two participants who had knowledge of the ultimatum game ($n_{Q7=17}$).

Analysis Methods We analyzed the results with Python data analysis libraries[15]. We run exploratory statistics, as well as hypothesis tests. First, we determine the differences between means and medians of game results and *niceness* evaluation of the two agent modes. For each of the hypotheses, we test the difference between two distributions using a Mann–Whitney U test[16]. We set the significance level α, as is common practice, to 0.05. To check for potential confounders, we calculate the Pearson correlation coefficient between demographic values and agent modes on the one side, and game outcomes and *niceness* scores on the other hand. In addition, we plot the *fairness* (ratio: coins received human / coins received agent) of the participant's game results. Furthermore, we summarize the participants' answers to the qualitative questions, which are also considered in the final, combined interpretation of the results.

21 persons participated in the study ($n_{init} = 21$). Two participants had detailed knowledge of the ultimatum game. We excluded the results of these participants from the data set ($n = 19$). The demographics of the participants are shown in Figure 2. The study participants are predominantly highly educated and "technical" (have a background in science and technology). People in their twenties and early thirties are overrepresented. This weakens the conclusions that can be drawn from the study, as less educated and less "technical" participants might have provided different results.

Table 2: Demographics of study participants

	Male	**Female**	
Gender	12	7	
	Bachelor	**Master**	**Ph.D. (or higher)**
Highest degree	2	12	5
	Yes	**No**	
STEM background	18	1	
Age (in years)	21, 25, 26 (2), 27, 28 (2), 29 (2), 30 (3), 31, 32 (2), 33, 40, 42, 62		

5.3 Results

Quantitative Analysis As can be seen in Table 3, notable differences in means and medians exist and are aligned with the assumptions that motivate the hypotheses. However, the differences are statistically not significant, as shown in Table 4. Considering the small sample size, no meaningful confidence interval can be determined[17]. When calculating matrix of correlations between demographics / agent mode and the different

[15] Data set and analysis code are available at `http://s.cs.umu.se/jo4bu3`.

[16] We choose the Mann–Whitney U test to avoid making assumptions regarding the distribution type of the game results and *niceness* score. However, considering the small sample size, strong, statistically significant results cannot be expected with any method.

[17] See, e.g.: [5].

Table 3: Results

	Not explained	Explained	Difference
Mean # rejections	1.67	1.5	0.17
Mean # coins, human	270.67	258.9	11.77
Mean # coins, agent	173.78	191.1	-17.32
Mean niceness score	3.11	3.6	-0.49
Median # rejections	2	1.5	0.5
Median # coins, human	283	260	23
Median # coins, agent	154	190.5	-36.5
Median niceness score	3	4	-1

Table 4: Hypothesis tests

Hypothesis	p-value
H_{a_0}	0.62
H_{b_0}	0.18
H_{c_0}	0.44
H_{d_0}	0.31

game result variables (and *niceness* scores), the *agent mode* does not stand out[18]. A noteworthy observation is the correlation of the participants' *gender* with the niceness score (3.71 for females and 3.17 for males, across agent modes).

When plotting the *fairness* (ratio: coins received human / coins received agent, see Figure 2), it is striking that one participant achieved an outstandingly high ratio of $11 : 1$. I.e., the human was likely deliberately and–in comparison to other participants–extraordinarily *unfair* to the robot. Considering this case an outlier and excluding it from the data set increases the difference between agent modes. Still, the difference is not significant (see Tables tables 5 and 6)[19]. When setting the thresholds for a *fair* game at a coins $agent : human$ ratio of $1 : 1.5$ and $1.5 : 1$, respectively, one can observe that in *explanations* mode, 70% (7) of the games are fair, whereas in *no explanations* mode, fair games amount for only 44.44% (4) of the played games.

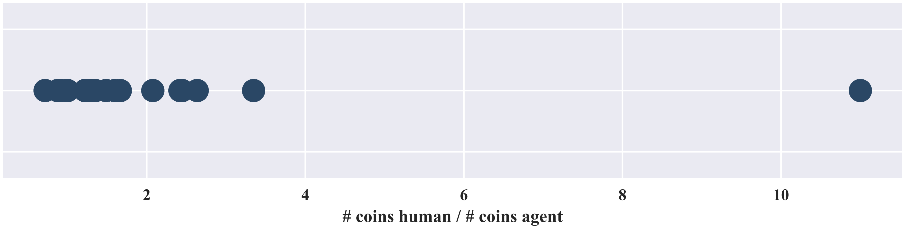

Fig. 2: Game fairness

[18] See the analysis notebook at `http://s.cs.umu.se/jo4bu3`.

[19] We consider the outlier detection and removal an interesting observation as part of the exploratory analysis that demonstrates the data set's sensitivity to a single extreme case. We concede this approach to outlier exclusion should be avoided when claiming statistical significance. Also, a multivariate analysis of variance (MANOVA) with the *game type* as the independent and *niceness*, *number of rejects*, and *number of coins received by the agent* as dependent variables did not yield a significant result.

Table 5: Results without outlier

	Not explained	Explained	Difference
Mean # Rejections	1.67	1.33	0.33
Mean # coins, human	270.67	257.11	13.56
Mean # coins, agent	173.78	209.56	-35.78
Mean niceness score	3.11	3.66	-0.56
Median # Rejections	2	1	1
Median # coins, human	283	250	33
Median # coins, agent	154	201	-47
Median niceness score	3	4	-1

Table 6: Hypothesis tests without outlier

Hypothesis	p-value
H_{a_0}	0.35
H_{b_0}	0.19
H_{c_0}	0.22
H_{d_0}	0.23

Qualitative Analysis

Interaction experience Generally, the participants noted that the robot had problems with processing their language, but was able to express itself clearly. It is noteworthy that two participants used the term *mechanical* to describe the robot in *no explanations* mode. In *explanations* mode, no such assessment was made.

Explanation evaluation Participant feedback on robot explanation was largely in line with the robot mode a given participant interacted with; i.e., participants who interacted with the robot in *explanations* mode found the explanations good (4 of the 7 participants who were asked **Q7**)) or somewhat good (3 out of 7)[20]. In contrast, most participants who interacted with the robot in *no explanations* mode typically noted that explanations were lacking/insufficient (7 of the 8 participants who were asked **Q7**). Constructive feedback one participant provided on the robot in *explanations* mode was to make explanations more *convincing* and *persuasive*. Another participant suggested the robot could *explain its strategy*.

5.4 Interpretation

Comparing the different explanation modes (explanation versus no explanation), there are *notable* differences in mean/median rejections, coins received by agent and human, and niceness score between the two modes. The differences reflect the initially stated assumptions that the agent that explains its proposals:

1. causes less proposal rejections;
2. receives more coins (while the human receives less);
3. is evaluated as *nicer* than the agent that does not explain its proposals.

However, as these differences are statistically not significant, empirically valid conclusions cannot be drawn. Considering the small size of the sample (number of study participants) and the language processing problems the participants reported when evaluating the interaction experience, it is recommendable to run the study on a larger scale

[20] Note that the sentiment was interpreted and aggregated by the researchers, based on the qualitative answers.

and improve the study design. In particular, the participant selection should be more diverse in their educational backgrounds, and the agent player should be more stable and less exotic; for example, a web-based chatbot with a simplistic graphical interface could be used to avoid the noise that was likely created by technical interaction difficulties and the humanoid robot's *novelty effect*.

6 Discussion

6.1 Sympathetic Actions of Learning Agents

Considering the increasing prevalence of (machine) learning-enabled agents, a relevant question is whether the concepts we presented above are of practical use when developing agents for human-computer interaction scenarios, or whether it is sufficient that the agent converges towards sympathetic behavior if deemed useful by the learning algorithm. One can argue that a powerful learning algorithm will enable an agent to adopt sympathetic behavior, even if its designers are ignorant of sympathetic actions as a viable option when creating the algorithm. However, the following two points can be made to support the usefulness of the provided goal types and explanations for sympathetic actions, even for learning agents:

- In practice, learning agents are incapable of executing "good" actions when they act in an environment about which they have not learned, yet. In the domain of recommender systems, which currently is at the forefront regarding the application of machine learning methods, the related challenge of providing recommendations to a new user, about whose behavior nothing has been learned yet, is typically referred to as the *cold start problem*[21]. One can assume that the concepts provided in this paper can be a first step towards informing the design of initial models that allow for better *cold starts* by enabling designers to create better environments and reward structures for learning agents that expect reciprocity from the humans they interact with.
- The provided concepts can facilitate an accurate understanding of the problem a to-be-designed learning agent needs to solve. As any machine learning model is a simplification of reality and as the temporal horizon a learning agent can possibly have is limited, the provided concepts can inform the trade-offs that need to be made when defining the meta-model of the agent and its environment, for example when determining rewards a reinforcement learning algorithm issues.

6.2 Limitations

This paper primarily provides a conceptual perspective on rational agents' goal types for and explanations of sympathetic actions, alongside with a preliminary human-agent interaction study. In the nature of the work's conceptual focus lies a set of limitations, the most important of which are listed below:

[21] See, for example: Bobadilla et al. [3]

- **The concepts lack empirical validation.**
  As stated in Subsection 5.4, the preliminary empirical study does not provide significant evidence for the impact of explanations on human-agent games. A more thorough empirical validation of the concepts is still to be conducted. In future studies, it would also be worth investigating to what extend *explainable* agents can mitigate comparably selfish attitudes humans have when interacting with artificial agents in contrast to human agents (i.e., a study shows that human offers in the ultimatum game are lower when playing with a machine instead of with another human [20]). However, we maintain that the introduced perspective is valuable, in that it considers existing empirical behavioral economics research; hence the provided concepts can already now inform the design of intelligent agents that are supposed to interact with humans.
- **The scope is limited to two-agent scenarios.**
  The presented work focuses on two-agent scenarios (one human, one computer). Games, with multiple humans and/or computer actors that play either in teams or *all-against-all* are beyond the scope of this work, although certainly of scientific relevance.
- **The focus is on simplistic scenarios and agents.**
  This paper provides simplistic descriptions of the core aspects of explainable sympathetic actions, with a focus on human-agent interaction scenarios. To facilitate real-life applicability, it is necessary to move towards employing the concepts in the context of complex autonomous agents. However, the perspective the concepts can provide on such agents is not explored in detail, although an application of the concepts in the design of learning agents is discussed in Subsection 6.1.

6.3 Future Work

To address the limitations of this work as described in the previous subsection, the following research can be conducted:

- **Empirically evaluate the introduced concepts by conducting human-computer interaction studies.**
  To thoroughly evaluate how effective the introduced concepts are in practice, human-computer interaction studies can be conducted. The presented preliminary study can inform future studies at a larger scale. We recommend conducting the study with a simplistic web-based agent instead of a humanoid robot. This *(i)* avoids setting the user focus on the novelty effect of the robot *per se*, *(ii)* prevents "noisy" data due to technology glitches that impact the interaction experience, and *(iii)* makes it easier to have more participants with more diverse backgrounds, as the study can then conveniently be conducted online. Studies of humanoid robots can complement the insights gained from web-based studies, for example by considering the impact of human-like facial expressions.
- **Consider scenarios with any number of agents.**
  It is worth exploring how sympathetic actions can affect interaction scenarios with more than two agents. The work can be related to behavioral economics, for example to an experiment conducted by Bornstein and Yaniv that investigates how

groups of humans behave when playing a group-based variant of the ultimatum game [4]; the results indicate that groups act more *rationally* in the sense of classical game theory.

- **Apply the concepts in the context of learning agents.**
  As discussed in Subsection 6.1, the proposed concepts can be applied when designing learning agents. In this context, additional human-agent interaction studies can be of value. For example, when playing a series of ultimatum games with a human, an agent can attempt to learn a behavior that maximizes its own return from the whole series of games by showing sympathetic (*fair*) behavior to incentivize the human opponent to accept the agents' offer and make generous offers themselves.Then, ethical questions arise that are worth exploring. E.g., as the sympathetic actions are seemingly sympathetic (or: *fair*), but *de facto* rational: is the human upon whose behalf the agent acts deceiving the human who plays the ultimatum game?

7 Conclusion

In this paper, we presented a taxonomy of goal types for rational autonomous agents to act sympathetically, i.e., to concede utility to other agents with the objective to achieve long-term goals that are not covered by the agent's utility function. The suggested combinations of sympathetic actions and different explanation types can be applied to agents that are supposed to be deployed to human-computer interaction scenarios, e.g., to help solve *cold start* challenges of learning agents in this context. The presented concepts and the presentation of a preliminary human-robot interaction study can pave the way for comprehensive empirical evaluations of the effectiveness of the proposed approach.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Bench-Capon, T., Atkinson, K., McBurney, P.: Altruism and agents: an argumentation based approach to designing agent decision mechanisms. In: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2. pp. 1073–1080. International Foundation for Autonomous Agents and Multiagent Systems (2009)
3. Bobadilla, J., Ortega, F., Hernando, A., Bernal, J.: A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems **26**, 225–238 (2012)
4. Bornstein, G., Yaniv, I.: Individual and group behavior in the ultimatum game: are groups more "rational" players? Experimental Economics **1**(1), 101–108 (1998)
5. Campbell, M.J., Gardner, M.J.: Statistics in medicine: Calculating confidence intervals for some non-parametric analyses. British medical journal (Clinical research ed.) **296**(6634), 1454 (1988)

6. Chandrasekaran, A., Yadav, D., Chattopadhyay, P., Prabhu, V., Parikh, D.: It takes two to tango: Towards theory of ai's mind. arXiv preprint arXiv:1704.00717 (2017)
7. Core, M.G., Lane, H.C., Van Lent, M., Gomboc, D., Solomon, S., Rosenberg, M.: Building explainable artificial intelligence systems. In: AAAI. pp. 1766–1773 (2006)
8. Dautenhahn, K.: The art of designing socially intelligent agents: Science, fiction, and the human in the loop. Applied artificial intelligence **12**(7-8), 573–617 (1998)
9. Defense Advanced Research Projects Agency (DARPA): Broad Agency Announcement - Explainable Artificial Intelligence (XAI). Tech. Rep. DARPA-BAA-16-53, Arlington, VA, USA (Aug 2016)
10. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
11. Fishburn, P.: Utility theory for decision making. Publications in operations research, Wiley (1970)
12. Güth, W., Schmittberger, R., Schwarze, B.: An experimental analysis of ultimatum bargaining. Journal of economic behavior & organization **3**(4), 367–388 (1982)
13. Harbers, M., Van den Bosch, K., Meyer, J.J.: Modeling agents with a theory of mind: Theory–theory versus simulation theory. Web Intelligence and Agent Systems: An International Journal **10**(3), 331–343 (2012)
14. Kahneman, D.: Maps of bounded rationality: Psychology for behavioral economics. American economic review **93**(5), 1449–1475 (2003)
15. Kampik, T., Nieves, J.C., Lindgren, H.: Towards empathic autonomous agents. In: EMAS 2018 (2018)
16. Kraut, R.: Altruism. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, spring 2018 edn. (2018)
17. Kulkarni, A., Chakraborti, T., Zha, Y., Vadlamudi, S.G., Zhang, Y., Kambhampati, S.: Explicable robot planning as minimizing distance from expected behavior. CoRR, abs/1611.05497 (2016)
18. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: AAAI. pp. 4762–4764 (2017)
19. Leslie, A.M.: Pretense and representation: The origins of "theory of mind.". Psychological review **94**(4), 412 (1987)
20. Melo, C.D., Marsella, S., Gratch, J.: People do not feel guilty about exploiting machines. ACM Trans. Comput.-Hum. Interact. **23**(2), 8:1–8:17 (May 2016). https://doi.org/10.1145/2890495, `http://doi.acm.org/10.1145/2890495`
21. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: Beware of inmates running the asylum. In: IJCAI-17 Workshop on Explainable AI (XAI). vol. 36 (2017)
22. Parsons, S., Wooldridge, M.: Game theory and decision theory in multi-agent systems. Autonomous Agents and Multi-Agent Systems **5**(3), 243–254 (2002)
23. Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.M.A., Botvinick, M.: Machine theory of mind. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4218–4227. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018)
24. Richardson, A., Rosenfeld, A.: A survey of interpretability and explainability in human-agent systems. XAI 2018 p. 137
25. Savarimuthu, B.T.R., Purvis, M., Purvis, M., Cranefield, S.: Social norm emergence in virtual agent societies. In: International Workshop on Declarative Agent Languages and Technologies. pp. 18–28. Springer (2008)
26. Thaler, R.H.: Anomalies: The ultimatum game. Journal of Economic Perspectives **2**(4), 195–206 (1988)