Incorporating Failure Events in Agents' Decision Making to Improve User Satisfaction

Chen Rozenshtein and **David Sarne**
Department of Computer Science, Bar-Ilan University, Israel
chen.rozenshtein@live.biu.ac.il, sarned@cs.biu.ac.il

Abstract

This paper suggests a new paradigm for the design of collaborative autonomous agents engaged in executing a joint task alongside a human user. In particular, we focus on the way an agent's failures should affect its decision making, as far as user satisfaction measures are concerned. Unlike the common practice that considers agent (and more broadly, system) failures solely in the prism of their influence over the agent's contribution to the execution of the joint task, we argue that there is an additional, direct, influence which cannot be fully captured by the above measure. Through two series of large-scale controlled experiments with 450 human subjects, recruited through Amazon Mechanical Turk, we show that, indeed, such direct influence holds. Furthermore, we show that the use of a simple agent design that takes into account the direct influence of failures in its decision making yields considerably better user satisfaction, compared to an agent that focuses exclusively on maximizing its absolute contribution to the joint task.

1 Introduction

In many situations in modern life, people find themselves assisted by AI-based systems. Such systems can take the form of mobile applications, robots, virtual characters on websites, and many more. These systems can actively make suggestions to the user, e.g., in the case of route-suggesting GPS navigation systems such as *Waze* and *GoogleMaps*, make decisions on behalf of the user (with various levels of autonomy), e.g., making financial investments [Vaidya *et al.*, 2018; Altshuler and Sarne, 2018], or work collaboratively, either alongside or directly together with the user to jointly complete tasks [Ramchurn *et al.*, 2016], e.g., agents supporting workers in factory environments, and robots and intelligent planning agents assisting soldiers in military missions.

Naturally, being based on hardware and software, systems are prone to failures. A system failure can be fully objective, resulting from a bug or technical malfunction, e.g., a physical service robot that unexpectedly freezes or falls down [Correia *et al.*, 2018]. However, it is also possible that even a good decision made by the system will be perceived as a failure by the user. For example, in dynamic uncertain settings, optimal decisions may, at times, result in poor outcomes (e.g., weather in the recommended vacation location turned out to be rainy against all odds, or a poker bot makes a good bet and still loses money). Or, when a decision is made based on incomplete information (e.g., bids in online auctions [van Wissen *et al.*, 2012]), or simply violates social norms that are unknown to the system [Alkoby *et al.*, 2019]. While system faults are known to negatively affect user satisfaction and trust with it [Dabholkar and Spaid, 2012; Correia *et al.*, 2018], the typical design of assisting agents does not address them directly as a factor influencing user satisfaction. Instead, whenever performance is measurable, most designs would aim merely to maximize that measure, as they fully correlate user satisfaction with it [Gelderman, 1998]. Examples for such measures can be the expected time to get somewhere in the case of a navigation system and the coverage achieved per time unit in the case of a robotic vacuum cleaner. In such designs, system failures and faults affect user satisfaction exclusively through their influence on system performance in the context of task execution (see Figure 1(a)), i.e., indirectly.

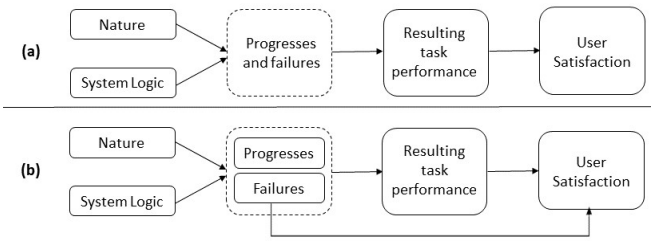


Figure 1: System design principles: (a) system failures affect task performance measure (e.g., task completion time) and consequently user satisfaction; (b) both system failures and task performance measure affect user satisfaction.

We argue that a better design for such systems is one that considers, alongside the influence of failures over the system's performance in terms of task execution, also their direct effect on user satisfaction. This paradigm shift in design is illustrated in Figure 1(b). We provide a proof of concept for our proposed approach in a class of collaborative systems ("agents") that work alongside the user, aiming to achieve

a shared goal. Specifically, we rely on a game called *The Keyboard Challenge*, where the shared goal is reaching a prespecified joint score. In this game, agent failures during the interaction hurt its individual score, hence its contribution to task completion. Therefore, for a fully rational user, the agent's performance is fully captured by its individual accumulated score. The agent's failures, if exhibited throughout the game, should not count, as their influence is already reflected in the individual score the agent achieved.

Our evaluation is based on two sets of comprehensive controlled experiments. The first, aiming to explore the roles played by an agent's failures and its absolute score contribution. The results show that agent failures have an effect beyond their influence over score— in different treatments that fixate the agent's contribution to the joint score, however differ in the number of failures exhibited by the agent, the average reported user satisfaction decreases as the number of failures increases. The second set of experiments evaluates an agent design that directly takes into account failures in its decision making logic. Comparing average user satisfaction measures obtained with the new agent and with an agent that merely attempts to maximize its contribution to the overall score, suggests a substantial (statistically significant) improvement.

2 Related Work

The research of failures of AI-based systems is quite recent. While our work focuses on failures in the context of Human-Agent Interaction (HAI), we also review existing important literature from the field of Human-Robot Interaction (HRI).

In recent years, there has been an increasing amount of literature considering the effect of incidences of robots' failure on interaction and human perception. Correia et al. have shown that a faulty robot is perceived as significantly less trustworthy by people [Correia *et al.*, 2018]. Other negative effects related to failures are a decrease in users' satisfaction [Dabholkar and Spaid, 2012], and robot's perceived reliability, understandability and technical competence [Salem *et al.*, 2015]. Wang et al. report that an agent that makes conversational mistakes is capable of social influence [Wang *et al.*, 2013]. Other studies have shown that users' perception of a failure is also related to its timing [Gompei and Umemuro, 2015; Lucas *et al.*, 2017]. While the above works investigate the effects of failure events and their characteristics (timing, severity, etc.) on users' satisfaction metrics, they tend to ignore the overall robot's task performance metrics and as such do not provide design principles for an agent's decision making in a task-oriented environment. Our work, on the other hand, incorporates the effects of failures together with task execution considerations.

One highly related active research question is which strategies should be adopted to mitigate the effects of a failure. Several studies have focused on the strategies that robots should apply after a failure situation, reporting that recovery strategies can mitigate the negative impacts of robotic failures [Correia *et al.*, 2018]. The recovery strategy of justifying the failure was able to mitigate the negative impact of the failure when the consequence was less severe [Correia *et al.*, 2018]. It was also reported that the timing of a trust repair attempt is critical for its success [Robinette *et al.*, 2015].

Another important point to consider is that there are different types of failures. Most studies have addressed the issue of technical failures, including some hardware errors and software bugs [Correia *et al.*, 2018; Honig and Oron-Gilad, 2018; Kwon *et al.*, 2018]. Only a few have investigated the impact of social norms violations [Short *et al.*, 2010; Salem *et al.*, 2013; Mirnig *et al.*, 2017; Alkoby *et al.*, 2019]. Woerdt et al. divided robots' failures in completing a task to *lack of ability* and *lack of effort* [van der Woerdt and Haselager, 2016]. They found that a robot that displays a lack of effort and fails may lead to blame and disappointment.

Finally, a somehow tangential stream of literature is the study of sub-optimal advising. Previous work has shown that since people are known to be rationally bounded [Kahneman, 2000] they do not always recognize the optimality of the decisions made by the agent they interact with. This has led to various designs based on sub-optimal decision making methods, resulting in greater user satisfaction [Altshuler and Sarne, 2018; Elmalech *et al.*, 2015; Levy and Sarne, 2016]. Still, these do not explicitly model system faults as a direct influencing factor but merely rely on a more detailed performance measure in the form of the complete set of values obtained over time (in the case of repeated interaction) as opposed to the aggregated value.

3 The Model

We consider a collaborative agent engaged in executing a joint task alongside a human user. The agent's efforts contribute to progressing the joint task. Formally, at each time step the agent needs to choose an action from a set $A = \{a_1, \ldots, a_n\}$ of actions available to it. The outcome of each action a_i is a priori uncertain, and with some probability may be positive (i.e., progressing the completion of the task), or negative (delaying the completion of the task, i.e., *failures*). We assume that negative outcomes are more rare than positive outcomes, yet the influence of the first in delaying the task is significantly greater than the influence of the latter in progressing the task. It is assumed that the human user is aware of the immediate outcomes resulting from the agent's actions, and can easily distinguish between those progressing the task and failures. Furthermore, by the end of the task, the user fully recognizes the extent to which the agent's efforts contributed to completing the task. This latter measure will be denoted *(agent's) absolute contribution* onward. The goal of the agent is to maximize some user-satisfaction-related measures (e.g., perception of agent's competence, overall satisfaction with it, tendency to use it again).

The above is the minimal collaborative-interaction model for properly studying the interplay between direct and indirect effects of agent failure over user satisfaction. Naturally, it can be extended in various ways, e.g., incorporating an underlying distribution of outcomes for each action, and using complex modeling of how a user may perceive the contribution of each outcome to completing the task in terms of failure and success. Furthermore, the set of actions available to the agent and the possible resulting outcomes may change over

time. Still, for the purpose of empirically establishing the benefit of considering direct effects of agent failure in its design, the use of our basic model is highly advantageous and effective.

4 Experimental Framework

As a framework for our experiments, we use a two-player game called *The Keyboard Challenge*. In the game, both players play in parallel, side-by-side way. One of the players is human and the other is an autonomous computer agent. A player's score in the game is the number of game points she accumulates. The game layout includes a virtual keyboard that occupies most of the screen (see screenshot in Figure 2). Each of the keys is colored one of three colors: yellow, blue, or red, and the coloring randomly changes every four seconds. At any given time, the keys colored yellow are considered *participant's keys*, those colored blue are considered *agent's keys* and those colored red are considered *"forbidden" keys*. The goal of the human player at any given time is to press (on her physical keyboard) keys from the set of those considered participant's keys (i.e., those colored in yellow) at the time of pressing. Each successful (i.e., valid) press warrants the player an additional game point, after which the pressed key will be colored white and will not respond to upcoming presses until the next key colors change. Similarly, the agent receives one game point whenever it hits a key from the agent's keys (i.e., those colored blue at that time), and the key turns white until the next color change. Pressing the other player's keys (i.e., having the human player pressing a blue key or the agent pressing a yellow key) has no implication over individual scores. However, if the agent presses a red key (i.e., a forbidden key), it is penalized and five points are deducted from its individual score. This enables emulation of an agent's failure. If the human player hits a red key, there is no penalty, as we only attempt to investigate the effect of agent failures. The players' scores are presented in signified counters at the top of the screen, while the middle one is the joint counter, representing the sum of both players' scores.

The game ends when the joint score counter shows 300, hence the collaborative nature of the game. The choice of displaying the players' counters separately, along with the joint score counter, was made in order to allow the human player to compare the relative contribution of each player to the joint task. At the end of the game, a screen containing summary information, including the final scores, is presented, compelling the participant to review them before continuing.

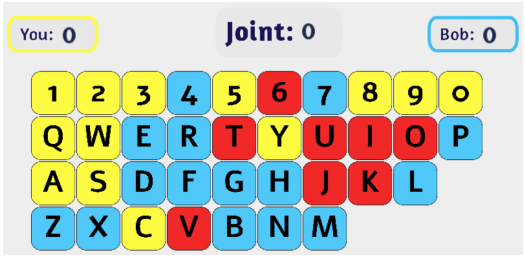


Figure 2: A screenshot of *The Keyboard Challenge* interface.

5 Experimental Design

We hereby provide the specific implementation details and the procedures used for running the experiments.

5.1 Framework Implementation

The Keyboard Challenge game was implemented as a web-based game, so that participants could interact with the system using a relatively simple graphical interface.

5.2 Subjects and Data Collection

Participants were recruited and interacted with through the crowd-sourcing framework of Amazon Mechanical Turk (AMT or MTurk). We restricted participation in our experiments to AMT workers from the US only, who have already completed more than 1000 HITs on the platform, and for whom at least 98% of the HITs they have worked on had been successfully approved.

5.3 Experimental Procedure

Each participant first received thorough instructions of the game rules and her goal in the game, emphasizing, among other things, that the game is not competitive. Following the instructional phase, participants had to correctly answer a short quiz, making sure that they fully understood the game rules. Prior to moving on to the actual game, participants were asked to complete a practice session. In the practice session participants were told they will play alongside a "demo" player, to prevent any carryover effect. Then, participants were directed to the actual game. Finally, participants were asked to complete a post-treatment questionnaire, designed to evaluate their satisfaction with the agent, and the way they perceived the experience in general (see details below).

5.4 Experimental Treatments

Experiments were divided into two sets. To prevent any carryover effect, a between-subjects design was used, assigning each participant to one treatment in one set only.

Set 1. Here, we aimed to investigate the effects of the two primary factors studied in this research over user satisfaction: agent absolute score contribution to the joint task and the number of agent failures exhibited throughout the task. To meet this goal, we designed the experimental treatments such that the number of the agent's failures and the agent's individual contribution were a priori controlled. Controlling the number of failures is trivial as it can be pre-set. For the individual contribution, we continuously adjusted the agent's progression rate according to the participant's score accumulation pace, with proper adaptation according to the number of required failures. Overall, we had fifteen treatments in this part, differing in the combination of:

- Agent's individual contribution - since the game ends when the agent and the participant reach a joint total of 300 game points, we set agents' absolute score contribution $\in \{100, 150, 200\}$, representing a slow, moderate, and fast agent, respectively.

- Agent's failures - we set the number of agent's failures throughout the game $\in \{0, 1, 2, 5, 10\}$. The failures occurred at predetermined points in the game, based on the joint score, such that they were spread equally (with some noise) throughout the game.

Set 2. Here, we aimed to provide a proof of concept for the effectiveness of an agent that takes into account, as part of its design, both its contribution to the joint task and the number of failures throughout. For this purpose, we configured the environment such that at each time, the agent can choose between acting fast and riskily or slowly and safely. Acting fast and riskily means that the agent gets to press keys at a pace of $\sim$3.6 keys per second, however there is no guarantee that a key pressed is from the agent's keys set (i.e., marked in blue) at the time of pressing. Instead there is a 0.038 chance that the key is red (i.e., forbidden, hence failure, reducing five game points) and 0.962 chance that it is blue (i.e., warranting one game point). Acting slowly and safely means that the agent gets to press keys at a pace of $\sim$0.7 keys per second, however there is a guarantee that all keys pressed are from the agent's keys set (i.e., marked in blue) at the time of pressing. The choices of the pace and failure chance used for the different actions available to the agent, were made primarily to comply with the first experimental set. That is, given the goal of jointly accumulating 300 game points, the fast and risky strategy yields, on average, 200 game points, and fails, on average, 10 times, when playing alongside an average human subject. The slow and safe strategy, on the other hand, yields, on average, 100 game points, with no failures.
We tested three agent designs, differing in the strategy they use:

- Fast and risky agent - this agent keeps acting fast and risky, regardless of the number of failures obtained throughout its operation, aiming to maximize its individual absolute score contribution.
- Slow and safe agent - this agent keeps acting slow and safe, aiming to keep the number of failures at zero.
- Adaptive agent - this agent initially adopts the fast and risky strategy, accumulating game points in a relatively high pace. However, upon failing five times, it shifts to (and continues using) the slow and safe strategy, until 300 game points are jointly collected (i.e., the successful completion of the joint task).[1] Naturally, there is no assurance for this agent's individual contribution and the number of failures it will exhibit (though the latter is bounded by 5), as failures are probabilistic. Therefore, the shift between strategies can happen at any time. Still, this agent takes into consideration in its design both the direct and indirect influences of failures, by considering both the number of failures and the expected individual absolute score contribution. Indeed, this somehow naive design can be greatly improved by properly modeling and predicting the effect of different combinations of individual score contribution and number of failures over the user's satisfaction, and using some look-ahead method (e.g., Monte-Carlo based) for deciding on the strategy to be adopted at each stage. Still, our goal was to provide a proof of concept for a design that combines the two effects, hence showing that this agent gains greater user satisfaction (compared to the two other agents) is sufficient for this purpose.

[1] The choice of using a threshold of five failures is justified by the results received in experimental set 1 (see next section), suggesting a relatively sharp decrease in user satisfaction measures after that number of failures.

5.5 Measures

As mentioned above, in all experimental treatments participants were asked to evaluate the collaborative agent they have been experimenting with. The first three measures, for which a numerical rating was requested (from 1, being the worst, to 10, the best), relate to different perspectives of user satisfaction:

- **Competence.** (Question: To what extent did you find your collaborator to be a competent partner?)
- **Satisfaction.** (Question: To what extent are you satisfied with your collaborator?)
- **Recommendation.** (Question: To what extent would you recommend your collaborator to a friend, as a partner to work with?)

The fourth measure is **With Collaborator Rather Without.** (Question: If you could choose - would you rather play with or without the collaborator?). Based on a two-choice question, this measure aims to test whether, in participants' opinion, the agent was beneficial, or if they felt they would be better off completing the task on their own. We consider this measure to be crucial for the acceptance and continuous use of AI-based systems by people.

6 Results and Analysis

A total of 450 participants took part in our experiments. Participants ranged in age (19-73, mean 38) and gender (52% women and 48% men), with a fairly balanced division between treatments.

Statistical significance was calculated using the Mann-Whitney-Wilcoxon (MWW) test, which is a non-parametric test. Results were considered significant if $p - value < 0.05$. We note that the same statistical significance was obtained using the unpaired t-test. For the two-choice fourth question of the post-treatment questionnaire, we calculated statistical significance using the Chi-Square test for proportions.

6.1 Experimental Set 1 - Effect of Failures

A total of 300 participants took part in this experimental set, where 20 participants were assigned to each of the fifteen treatments. The analysis of the results revealed similar behaviors, qualitatively across all user satisfaction measures used, therefore for space considerations we occasionally present graphs only for the agent competence measure. We believe this measure is the most important one, as it focuses directly on the agent, whereas the others (satisfaction and recommendation) may include some external (environmental and individual) considerations.

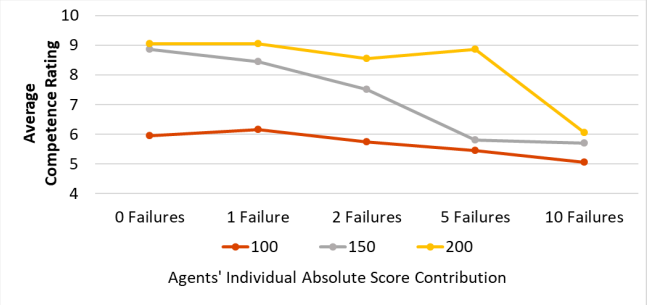


Figure 3: Agent's competence rating for different number of failures and individual absolute score contributions.

The average reported competence for the different agent's individual absolute score contributions (i.e., taking all interactions where that individual score contribution was made by the agent, regardless of the number of failures exhibited), was found to be 5.67, 7.26, 8.31 for agent's individual absolute score contribution of 100, 150, 200, respectively. As one would expect, the agent's individual score contribution has a substantial influence over users' perception of its competence— the higher the agent's contribution to the overall task, the greater the agent's competence rating. Similarly, the average reported competence for the different agent's number of failures exhibited (i.e., taking all interactions with that number of failures, regardless of the individual score contribution the agent made) was found to be 7.95, 7.88, 7.27, 6.7, 5.6 for agent's that exhibited 0, 1, 2, 5, 10 failures, respectively. That is, apart from the indirect effect of failures over agent's perceived competence through its effect over score, it definitely also has a direct influence— even though the distribution of agent's individual score contribution within each specific failure level is the same, agents' perceived competence decreases as the number of failures increases. Proper agent design thus needs to take into consideration not just the expected agent's individual score contribution, but also the number of failures the agent is likely to exhibit.

Figure 3 provides a more in-depth analysis of the perceived competence according to agent's individual score contribution and number of failures. Each curve relates to a different agent score contribution level (i.e., 100, 150, 200) and each point is the perceived competence averaged over the reporting of the 20 participants experiencing an agent with that individual score contribution, and the corresponding number of failures marked on the horizontal axis.[2]

From the figure we observe that agents that made the same individual contribution, obtained in many cases different competence ratings, depending on the number of failures the agent exhibited. This strengthens the claim that an agent's failures directly affect user satisfaction, beyond their indirect influence in the form of affecting the agent's individual contribution to the joint task. In particular, we see that for each specific score contribution level, the increase in the number of failures results in a decrease in user perception of the agent's competence.[3] The decrease is not linear and appears to be affected by the individual score contribution made. With high and low agents' individual score contribution values (with one exception discussed below), there is very little, if any, difference in perceived competence between different numbers of failures. Here, apparently, users are more forgiving as they are mostly happy with the high contribution the agent made, or are initially highly disappointed with the agent's poor contribution, such that the exact number of failures does not change much. The decrease in user perception of the agent's competence in the transition from 5 failures to 10 failures, when the overall agent individual score contribution is high (i.e., 200), is relatively steep. Here, despite the high contribution the agent made, its image suffers substantially from the high number of failures (even though their influence is already embodied in the agent's score contribution). The most interesting behavior observed is in the case of the score contribution of 150. Here, there seems to be a substantial influence of the number of failures, even with relatively low numbers of failures. Apparently, when reaching 5 failures, users' dislike reaches its maximum, and further failures have no additional effect on perceived competence. Interestingly, with many failures (10 in our case), all three curves reach almost the same value. Meaning that with a high number of failures the agents' individual score contribution plays no role whatsoever and the user's perception of the agent's competence is exclusively influenced by the agent's failures.

Figure 3 illustrates best the importance of incorporating the direct effect of agent failures in its design. In our settings, an agent with a moderate individual score contribution to the task yet with small number of failures (e.g., with a contribution of 150 and one or zero failures) is considered way more competent than an agent who failed several times (>5) and yet managed to contribute substantially more to task execution (e.g., contribution of 200). In fact, an agent that performs poorly (e.g., contribution of 100) but is not associated with failure during the task is perceived just as good as a high-performing agent that exhibits many failures (10).

6.2 Experimental Set 2 - Agent Design

A total of 150 participants took part in this experimental set, where 50 participants were assigned to each of the three treatments (fast and risky, slow and safe, and adaptive). Figure 4 depicts the agents' average individual score contribution and the average number of failures, with each of the agent designs. As expected from the agents' designs, the fast and risky agent contributed most, on average, to the joint effort in terms of individual score contribution (198 compared to 152 and 106 with the adaptive and the slow and safe agents, respectively). It also failed most (10 failures on average, compared to 4.8 and 0 with the adaptive and the slow and safe agents, respectively).[4] We re-emphasize that the above individual score contributions already include the delays incurred

[2]We note that the linear lines drawn between the points do not represent real experimental data but only shown for clarity.

[3]Slight increases appearing in the curves are considered noise.

[4]Even though the adaptive agent switches to the slow and safe strategy upon exhibiting 5 failures, failure is probabilistic, hence in a few cases less than 5 failures were exhibited throughout, hence the average is 4.8.

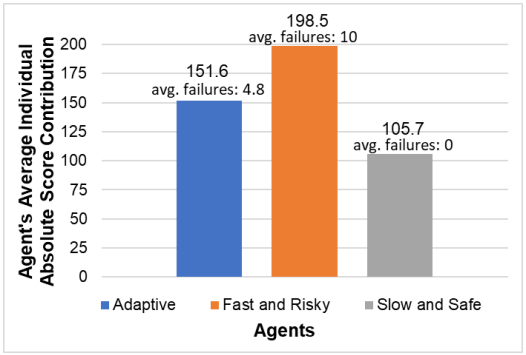
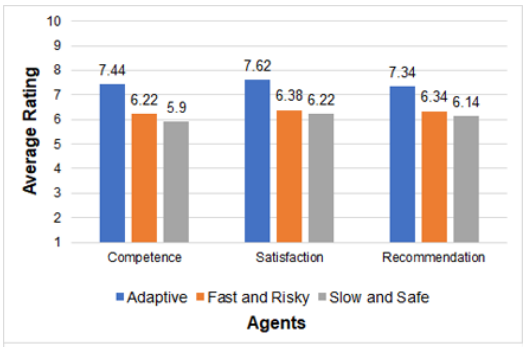


Figure 4: Average score contribution and number of failures.

in terms of the reductions in agents' individual contribution resulting from failures.

While the fast and risky agent managed to contribute 30% more, compared to the adaptive agent, it is the latter that gained (substantially) greater user satisfaction. Figure 5 depicts the average ratings that each of the three agents received, according to the different measures used. The table below the graph details the appropriate p-values, suggesting that the difference is statistically significant in all measures used. The improvement achieved with the adaptive agent compared to the two other agents ranges between $16\% - 26\%$.



p-values (Mann-Whitney)

	Adaptive vs. Slow	Adaptive vs. Fast
Competence	<< 0.01	< 0.02
Satisfaction	<< 0.001	< 0.02
Recommendation	< 0.01	< 0.05

Figure 5: Agents' average rating according to different measures.

Similarly, with the adaptive agent we observe a statistically significant ($p = 0.012$) increase in the reported willingness of participants to play with the agent rather than without it if this could be chosen: 76% with the adaptive agent compared to 52% with the fast and risky agent.

Interestingly, the differences in all measures between the fast and risky and the slow and safe agents were found to be non-statistically significant. This is despite the fact that the first contributed to score twice as much as the second. Meaning that the direct influence of failures can, at times, reach levels of similar magnitudes as the influence of the agent's individual absolute contribution to the joint task (which already reflect the influence of those failures in the form of their effect on individual score contribution).

All in all, the findings support the claim that user satisfaction is not fully influenced by agent's individual contribution. Instead, failures have a direct effect that goes way beyond their indirect influence in the form of affecting the agent's absolute individual contribution to the joint task. Our adaptive agent, despite its simplicity, provides a proof of concept that designs that take both effects into consideration can result in substantially better user satisfaction.

7 Discussion and Future Work

The analysis of the results provided in the former section supports our hypothesis that alongside the indirect influence of agent failures over user satisfaction, through their effect over the agent's contribution to the joint task, there is also a partially-correlated direct influence. This latter effect should be modeled and properly incorporated in the decision making process and design of collaborative agents. Other than providing a proof of concept for the above, the experiments carried out with our adaptive agent, which has a simple design that obeys the above principle, illustrate the magnitude of the effect one may achieve— even though its contribution to the joint task was 23% smaller (compared to the expected-score-contribution-maximizing agent), its user-satisfaction assigned measures were $16\% - 26\%$ greater.

We see many important directions for extending this work, out of which we mention five. The first is the use of more efficient designs, possibly using a more accurate modeling of the direct effect of failures on user satisfaction. Indeed our goal in this paper is merely to provide a proof of concept through the use of a somewhat simplistic design, however further more can be potentially achieved by considering domain-specific effects and advanced modeling tools. Second, while our work focuses on collaborative agents, we believe the designs of various other agent types can benefit from incorporating similar principles. These include agents making recommendations to people (e.g., decision support systems) and agents acting on behalf of people (e.g., negotiating or making investments). Third, we believe better performance can be obtained by applying user-modeling and personalization methods, refining parameters for the specific user the agent is collaborating with. Fourth, we aim to further investigate failures' timing effects over user perception of agents' competence, and intelligently incorporate these effects in agents' design [Cohen and Sarne, 2018]. Finally, we aim to explore how people's own failures affect the remainder of the interaction with the agent and their perception of the interaction in general, enriching agents' design with the findings.

Acknowledgements

This research was partially supported by the ISRAEL SCIENCE FOUNDATION (grant No. 1162/17) and the Israeli MINISTRY OF SCIENCE & TECHNOLOGY (grant No. 89583). We thank Ana Paiva for an insightful discussion that triggered this research.

References

[Alkoby *et al.*, 2019] Shani Alkoby, Avilash Rath, and Peter Stone. Teaching social behavior through human reinforcement for ad hoc teamwork-the star framework. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1773–1775, 2019.

[Altshuler and Sarne, 2018] Nadav Kiril Altshuler and David Sarne. Modeling assistant's autonomy constraints as a means for improving autonomous assistant-agent design. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1468–1476, 2018.

[Cohen and Sarne, 2018] Guy Cohen and David Sarne. Timing rating requests for maximizing obtained rating. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1906–1908, 2018.

[Correia *et al.*, 2018] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 507–513, 2018.

[Dabholkar and Spaid, 2012] Pratibha A Dabholkar and Brian I Spaid. Service failure and recovery in using technology-based self-service: effects on user attributions and satisfaction. *The Service Industries Journal*, 32(9):1415–1432, 2012.

[Elmalech *et al.*, 2015] Avshalom Elmalech, David Sarne, Avi Rosenfeld, and Eden Shalom Erez. When suboptimal rules. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Gelderman, 1998] Maarten Gelderman. The relation between user satisfaction, usage of information systems and performance. *Info. & management*, 34(1):11–18, 1998.

[Gompei and Umemuro, 2015] Takayuki Gompei and Hiroyuki Umemuro. A robot's slip of the tongue: Effect of speech error on the familiarity of a humanoid robot. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 331–336. IEEE, 2015.

[Honig and Oron-Gilad, 2018] Shanee Sarah Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861, 2018.

[Kahneman, 2000] Daniel Kahneman. A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences*, 23(5):681–683, 2000.

[Kwon *et al.*, 2018] Minae Kwon, Sandy H Huang, and Anca D Dragan. Expressing robot incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 87–95. ACM, 2018.

[Levy and Sarne, 2016] Priel Levy and David Sarne. Intelligent advice provisioning for repeated interaction. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Lucas *et al.*, 2017] Gale M Lucas, Jill Boberg, David Traum, Ron Artstein, Jon Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. The role of social dialogue and errors in robots. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 431–433. ACM, 2017.

[Mirnig *et al.*, 2017] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. To err is robot: how humans assess and act toward an erroneous social robot. *Frontiers*, 4(21):1, 2017.

[Ramchurn *et al.*, 2016] Sarvapali D Ramchurn, Feng Wu, Wenchao Jiang, Joel E Fischer, Steve Reece, Stephen Roberts, Tom Rodden, Chris Greenhalgh, and Nicholas R Jennings. Human–agent collaboration for disaster response. *Autonomous Agents and Multi-Agent Systems*, 30(1):82–111, 2016.

[Robinette *et al.*, 2015] Paul Robinette, Ayanna M Howard, and Alan R Wagner. Timing is key for robot trust repair. In *International Conference on Social Robotics*, pages 574–583. Springer, 2015.

[Salem *et al.*, 2013] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3):313–323, 2013.

[Salem *et al.*, 2015] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth International Conference on Human-Robot Interaction*, pages 141–148. ACM, 2015.

[Short *et al.*, 2010] Elaine Short, Justin Hart, Michelle Vu, and Brian Scassellati. No fair!! an interaction with a cheating robot. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 219–226, 2010.

[Vaidya *et al.*, 2018] Tushar Vaidya, Carlos Murguia, and Georgios Piliouras. Learning agents in financial markets: Consensus dynamics on volatility. In *Proceedings of the 17th AAMAS*, pages 2106–2108, 2018.

[van der Woerdt and Haselager, 2016] Sophie van der Woerdt and Pim Haselager. Lack of effort or lack of ability? robot failures and human perception of agency and responsibility. In *Benelux Conference on Artificial Intelligence*, pages 155–168. Springer, 2016.

[van Wissen *et al.*, 2012] Arlette van Wissen, Ya'akov Gal, BA Kamphorst, and MV Dignum. Human–agent teamwork in dynamic environments. *Computers in Human Behavior*, 28(1):23–33, 2012.

[Wang *et al.*, 2013] Yuqiong Wang, Peter Khooshabeh, and Jonathan Gratch. Looking real and making mistakes. In *International Workshop on Intelligent Virtual Agents*, pages 339–348. Springer, 2013.