When Computers Should Say They Are Sorry:

Adaptive Versus Unconditional Apologeticness

Tinka Zorge
Graduation Thesis, September 2019
Media Technology MSc program, Leiden University
Supervisors: Catholijn Jonker and Myrthe Tielman
tinkazorge@gmail.com

Affective cues such as apologetic messages are an important tool to influence user experiences. For this study we designed an experiment to determine whether a computer teammate that takes responsibility and apologizes no matter who made the mistake is more frustrating for the user than a computer teammate who does this only when the user perceives them at fault for the mistake. We found that the users start out satisfied with the unconditional apologies but become more frustrated with them over the course of the experiment, ending up more satisfied with the teammate that apologizes only when perceived by the user as being the one at fault. The findings of the study contribute to determining preferred ways of handling crisis situations, meaning a situation in which a mistake is made that must be handled, with or without apologies in the fields of UX and HCI. *(Satisfaction, frustration, apology, usability, responsibility, interaction, user response, UX, HCI)*

I. Introduction

Over the years, countless researchers have been analyzing why and to what effect people are influenced by apologies from other people. The mechanism behind the apology can be interpreted in many ways, but its effect itself is undeniable when talking about human on human interaction. When it comes to computers, users also tend to rate apologetic systems as more satisfactory [Park, 2012]. However, in many different instances, apologies between people can fail to reach the desired effect [Wohl, 2011] [Zheng, 2016] [Gardner 2018] [Chiles, 2015]. Clearly, some boundaries exist on the effectiveness of the apology as a tool for social satisfaction between people. In this study, we are looking at those same boundaries of the interaction between humans and computer systems.

Our research question therefore is: when mistakes are made, do users prefer their computer systems to be apologetic regardless who they (the users) perceive as being at fault or is there a point where unconditional apologeticness decreases their satisfaction compared to selective apologeticness?

Based on situations in which apologies between humans have proven to have failed, which will be further explained below, we hypothesized that:

- The user would first be more satisfied by the apologies of the teammate that sends them unconditionally (UT) than those of the adaptive teammate that sends them selectively dependent on the situation (AT), but would at a certain point start regarding UT's continuous apologies as insincere, leading them to be less satisfied with the UT than the AT. There would be a turning point after a certain number of mistakes where the responses to the mistakes by the UT would start to yield more frustration from the user than those made by the AT.
- Overall, the users would choose the AT to play another round with in the future.

Our findings may contribute to determining preferred ways of handling crisis situations with or without apologies in the fields of UX and HCI.

II. Apology effectiveness between people

An apology can be regarded as an act that is meaningful for the victim and difficult for the transgressor because the one who apologizes must acknowledge that they "failed to live up to values like sensitivity, thoughtfulness, fairness, and honesty" [Lazare, 1995]. Here, an apology is regarded as an act that is socially costly and therefore might yield forgiveness in the other party. Hewitt describes how he regards the effectiveness of the apology as rather mysterious, since it is demanded and given so frequently in our social domain. It seems to have a great effect on social balances but it is unclear why: it allows people to move on after an incident as if nothing happened [1992]. On a larger scale, apologies are an important moral tool both within a society and between societies; they are usually regarded as "the right thing to do" after harm has been done. A full apology, in which remorse is expressed and responsibility is claimed, can move the perpetrator into a new moral space: from wrongdoer to person whose wrongs have been acknowledged and perhaps forgiven [Levinas, qtd in Wohl, 2011]. The wrongdoer also acknowledges the boundaries of acceptable behavior by apologizing [Wohl, 2011]. The morality of the apology has the high ground even over money: Abeler even uncovered in an experiment that offering a "cheap talk" apology after an unsatisfactory purchase relieved customer frustration more than a monetary compensation [Abler, 2010]. All in all, the effectiveness of the apology is undeniable.

III. Apology effectiveness between people and systems

It seems that an important factor that is called on to explain the effectiveness of the apology is the costliness of the apology message for the transgressor. We could reason that in working with systems, this mechanism might not work the same way, since we know that a system that apologizes is not actually sacrificing any social capital: it is simply programmed to apologize after certain incidents. On the other hand, we could

argue that setting up an apologetic system, rather than a neutral system or a system that blames the user, proves that the makers of the system decided to take the responsibility. In an experiment where an agressive, a neutral and an apologetic system were compared, the users favored the apologetic system [Park, 2012.

IV. When apologies might fail

However, there is also evidence that apologies between humans do not always have the desired effect. There can be many different reasons why this might happen. An instance in which apologies can fail is if the timing is off, such as when the apology is offered before or during the event that it is referring to. An example is when a person apologizes for not being able to make it to an appointment while the appointment has not happened yet. This type of apology might make a bad situation worse [Skarlicki, 2004]. Assuming that the main goal of apologies is to repair the relationship between perpetrator and victim, apologies can also fail when the victim feels like the appropriate response to the event is not forgiveness but demanding justice, whether an apology was given or not [Wohl, 2011]. In this light, forgiveness would be considered a weak response to the apology and therefore the apology does not relieve frustration but does the opposite. Another reason why the apology might not be accepted is because the relationship between the humans involved does not allow the apology to be effective, such as when a power imbalance exists that makes the apology of the aggressor to look cynical to the victim. The apology is then considered insincere [Zheng, 2016]. When we are talking about insincere or "empty" apologies, what is meant is an expression of regret and an acceptance of responsibility that has no emotional truthfulness. Gardner writes:

"Stock apologetic phrases such as "I regret what I did" and "I'm sorry for what I did" are not always used to express the speaker's emotions and—more to the point—they are not always understood to do so. Sometimes, they are understood on both sides as performative utterances by which the utter accepts responsibility, without any element of emotional report or emotional expression. Such an apology is sometimes disparaged as meaningless or empty" [2018].

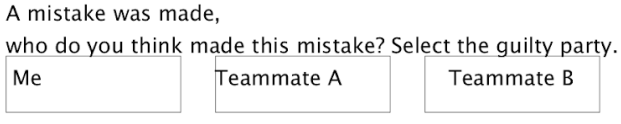
It is for this reason that over-apologizing can counterbalance the positive effects of an apology [Chiles, 2015]. After all, apologizing when the apology is not warranted might lead people to believe there is less emotional report behind your apologies, and therefore make them less effective. Even though the research by Park suggested people prefer their computer systems to be apologetic, we might expect that there are sharp boundaries on the amount and types of apologies humans accept from computer systems. In this study, we will not be trying out multiple types of apologies, but focus on one type of apology message that will be given to users in crisis situations under certain conditions: one teammate will apologize to the user unconditionally, and the other will only apologize only when considered at fault by the user. In short, we will be looking at the boundaries of unconditional apologies on relieving user frustration.

V. Method

To find the answers to this question we designed a within-subject lab study. In the experiment we had participants complete a memory task together with two computer teammates. The participants were presented with a little clip showing six household objects in a certain order. Afterwards they were asked to put the images in the correct order, as displayed below.

They were then shown subsequently what order their teammates chose. The order of each teammate was displayed on a seperate screen and the user was not able to move back to the previous screen. This was done purposely to create some ambiguity about who was at fault for possible mistakes, in an attempt to mirror the ambiguity about responsibility that is present in real-world situations.

After studying the orders given by the two teammates, the system checked the answers of the user and the teammates. If no mistake was made, the team of the user and their two teammates received a point. If a mistake was made, the user was prompted to decide who was responsible for the mistake: themselves, teammate A or teammate B.



When a mistake was made the unconditional teammate (UT) always apologized, no matter who the user perceived as

being at fault. The adaptive teammate (AT) only apologized when it was at fault in the eyes of the user. If the user chose the AT as the guilty party, the user would therefore receive two apologies. After getting one or more apologies, the user was prompted to answer some feedback questions to measure their frustration.

A. Pre-pilot

The goal of the pre-pilot was to run the experiment to find out how many mistakes users would have to make to start approaching a turning point where they felt more frustration than satisfaction from the apologies of the UT. Of course the existence of the turning point still had to be proven, but the pre-pilot was an indication if it could be found on a small number of participants. The experiment was set on a rather difficult level, as none of the four users managed to score a point. There was no clear turning point, but for some of the participants it seemed that around the fifth mistake the satisfaction of AT's responses topped the satisfaction of the responses by the UT.

B. Pilot

Knowing that we would want to get the participants to at least five mistakes to start approaching the turning point as suggested by the pre-pilot, the pilot was used to play around with the difficulty of the game and the number of rounds to see how to get the participants to at least five mistakes but preferably not all mistakes to keep them from losing all hope and with it, interest. We ended doing ten rounds to leave the participants enough chances to make mistakes as well as scoring some points. The difficulty of the game for the participants was mostly determined by the number of pictures they were shown and the speed with which the pictures would appear and disappear. Six pictures shown rather quickly back to back proved difficult enough for most participants to make at least five mistakes over the course of ten rounds.

C. Effect size and participants

We determined that the approximated effect size of our experiment should be n = 30. Koeman notes that in the 678 Human Computer Interaction/ User Experience (HCI/UX) laboratory experiments she looked at the median number of subjects for a lab study, which is "defined as a study that takes place in a controlled environment, often consisting of short sessions where participants carry out defined tasks in the presence of the researcher" is 23.4 [2018]. This study also concerns a HCI/UX laboratory experiment and contains a simple comparison of two within-participants conditions, we expect the number of participants necessary to see an effect to be similar to this average. To ensure effective results, we went up to 30 participants, which is also the number of participants Park used in his comparable study [Park, 2012]. The participants consisted of 9 males and 21 females between the ages of 18 and 56, the majority being between 18 and 26.

D. Measures

After each round of the task, we measured the user satisfaction about the response of each teammate. After all the rounds were over, we asked the user to choose a teammate to play another round with in the future. These are the variables we used to determine whether the users preferred the UT or the AT. To determine the satisfaction of the user about the response after each round, we used a scale from -4 (very frustrated) to +4 (very satisfied), with 0 as the neutral choice in the middle. The reason we chose a scale of -4 to 4 is because we would have a negative point, a positive point and perfectly neutral in the middle. On a scale from 1 to 10 or 0 to 9 the experience slowly climbs, which is not the intended progression.

Next to these, we controlled for a number of variables to be able to exclude their effects from our variables of interest. Our number of participants might be too low to draw any significant conclusions from these control variables, but strong connections might be interesting points of further study. First of all, we asked participants about their gender, age and nation of birth. We expected that different age groups and genders might have been raised with different perceptions of how computer systems and humans should behave towards each other. We reasoned that cultural background might be of interest in this study since there is evidence of significant cultural differences in the handling of apologies and responsibility. For instance, in individual agency-cultures apologies are seen as mechanisms for assigning blame and re-establishing personal credibility, while in collective-agency cultures, such as Japan, apologies are seen as utterances of true remorse [Maddux, 2011]. Countries that exhibit a more collective oriented culture as opposed to a more individualistic one are for instance located in West-Africa and East-Asia, while more individualistic countries can be found in North America, Western Europe and Australia [Kito, 2017].

Another variable we controlled for is how difficult the users perceived the task to be for the computer. We expected it might cause extra frustration if the user was convinced the task was easy for the computer, reasoning that it is making its mistakes purposely to fit the experiment, so we told the participants that for the computer this was an image recognition task. It is relatively well known that subjective visual image recognition is hard for computer systems [Buhmann, 1999]. Since we cannot be completely certain that every user is aware that image recognition is a difficult task for a computer, we asked a control question after the instructions had been made clear to see how hard the user expected the task to be for their teammates.

One control variable that turned out to be necessary came to our attention in the pilot, when two participants preferred UT over AT because they felt AT did not send enough messages in general. They preferred more communication in their computer systems overall. We therefore added a control question, asking participants by the end of the experiment how they felt about the amount of questions asked by AT and UT.

A last variable we chose to examine was the perceived performance of AT and UT throughout the rounds. The teammates had the exact same level of skill, but we wondered whether increased apologies would make the teammate seem more or less capable to the user. We therefore

asked the user what they thought of the performance of At and UT after each round, on a scale from -4 to +4 as well.

E. Procedure

Before the experiment started, we gave the participants clear instructions on a printed page with some imagery of the key elements they needed to understand. After reading those instructions carefully, they were asked to sign a consent form, explaining what would happen to the data and affirming that they could stop at any time. After signing the consent form, the participants were shown the first screen, on which another set of instructions was shown, explaining essentially the same rules but in different words and showing more visual examples of how the task was to be completed. By the end of the instructions, the user was informed that they would be able to obtain a special prize if they completed the task with the most points out of all the participants.

After the instructions, the participants played ten rounds of a memory task, as explained above. First, the user was presented with six pictures that followed one another quickly in a certain order, each one disappearing before the next one showed up. Then the user was presented with a row of six pictures that had the same items on them but were different images, which they had to put in the right order by numbering them. The user could then click continue and was shown the order that the adaptive teammate (AT) filled out, followed by the order that the unconditional teammate (UT) filled out. The users got all the time they needed to study their teammates' answers and come to their own conclusions about whose answer was different from their own answer. They clicked continue again and it was revealed whether a mistake was made.

The users now knew a mistake was made, and were asked who they thought made the mistake. They choose themselves or either one or both of the teammates. This is where the teammates' behavior started to differ. After the choosing of the culprit, the user got a screen informing that feedback would follow, and here they could get messages from the teammates, which were signified by a short message sound and an envelope icon on the middle of the screen. UT always sent an apology if a mistake was made by anyone, and AT only sent an apology when chosen as a guilty party. After receiving the message, the user went on to the feedback screens. If neither the user nor the teammates made a mistake, the user got to the feedback screens right away.

On the first feedback moment, the users were asked to choose on a scale of -4 (very frustrated) to 4 (very satisfied) how they felt about the performance of teammate AT and UT. For the second feedback question the users were asked how satisfied they felt about the response of teammate AT and UT. After the feedback questions were filled out, the users were informed about the number of points they had gotten so far and they could go on to the next round, where they would play the memory game again and give the feedback again.

After the users had done the ten rounds, they were asked to fill out their gender, age, nationality and email address. There was one final round of feedback: the user was asked how they felt about the amount of messages they got overall from AT and UT (on a scale from -4 to 4 again). They are then asked which teammate they would rather play with in the future, AT or UT. For the last question the participants are prompted to judge how hard they thought the task was for the computer on a scale from 1 to 10.

Lastly, the participants received a debrief document explaining what the experiment was really about and another one containing the contact information of the main researcher.

F. Task

Phase 1 of the experiment: introduction and consent

A double set of instructions was put in place to ensure that participants who did not fully understand the experiment when reading the first set of instructions would get another chance to fully grasp how it worked. We needed to be sure the participants understood everything properly, to make it unlikely for them to blame their mistakes on the design of the system instead of themselves, since we expected the most interesting feedback to arise from a situation in which they took responsibility for mistakes.

By promising the users an extra prize if they were the most successful at the experiment and awarding them points if they succeeded, we tried to make the task more important to the users. It has been suggested that there is a significant relation between frustration and the importance of the task [Ceaparu 2004].

Phase 2 of the experiment: memory game and feedback

A memory-orientated task was chosen because we expected that the participants might be more likely to blame themselves if the mistakes they made were made in their own head, rather than in a game where the mistake could be blamed on the design of the system. The initial idea was to have them complete a task that had to do with speed or precision, but we expected the possibility of the user blaming the design of the system instead of their own mistakes to be higher than in a memory related task.

Another aspect that argues in favor of the experiment we chose is that we intended to give the user the impression that the game was also hard for the computer. This had little to do with memory and more with the fact that we decided to have the user remember pictures and slightly mislead them into thinking that the computer had its own task in the matter, namely image recognition. We did not want the user to be more frustrated with their teammate simply because they felt the task was much easier for the computer and the mistakes were made by the computer on purpose: it has been established that when harm is caused purposely it leads to more frustration for others than when it is caused accidentally [Ames and Fiske, 2013]. By claiming the computer teammates had to work hard as well, we hoped to elevate both parties to a more equal level.

The amount of six items was chosen based on evidence that people can repeat no more than a random order of seven items [Miller, 1956]. According to Cowan [2010], it is even less: only three to five items for young adults when no tricks can be used by the subject. In our experiment, the time span

between remembering and reproducing was very short and the items familiar, so we started out with eight objects in the pre-pilot to ensure the users would make enough mistakes. We reduced this number to six after observing the high number of mistakes made. We used neutral images of six different objects, all used in Tengs memory experiment [Teng, 2010].

The computer teammates showed the correct answers—and therefore the same answers—by far most of the time. They each made one mistake over the course of ten rounds, simply to ensure that users that were unusually good at the memory game and did not make mistakes themselves also experienced both teammates apologizing. However, this was hardly necessary, as explained in the description of the pre-pilot.

After it was revealed to the users whether a mistake was made, the user could not know for sure who was at fault. This ambiguity was created purposely to mimic real world crisis situations in computer systems: often it is not clear where the responsibility of the error lies. When designing for UX it does not matter where the true responsibility lies either—the only thing that matters is the user perception of this and how they wish the system would respond.

An interesting and vital part of setting up the experiment was the design of the apology message. A lot of research has been done on the elements that a successful apology message should contain, and as the overview of research by Bentley shows, a consensus has not exactly been reached [2017]. There are of course elements that overlap in most studies, which are the elements that we used to craft an effective but not exaggerated apology message. Pace et al. constructed an experiment evaluating apology messages and concluded that expressing remorse and accepting responsibility made a message more effective, so these are the main elements that the message contains. Just stating "I'm sorry" without a clear taking of responsibility is possibly not enough to be effective. Wohl states: "Simply saying 'I'm sorry' may operate to smooth relations, but without taking responsibility for illegitimately harming another, the apology is hollow and unlikely to create lasting relationship change" [Wohl, 2011]. Therefore, because expressing remorse and accepting responsibility are the two most important elements of most of the studies, these are the ones that we conveyed in the apology message send to the user.

G. Ethical approval

Since this experiment contained deceiving of the participants (they were told the task was difficult for the computer due to image recognition even though this was not true), we obtained written approval from the ethics committee of Delft University prior to the gathering of data.

VI. Results

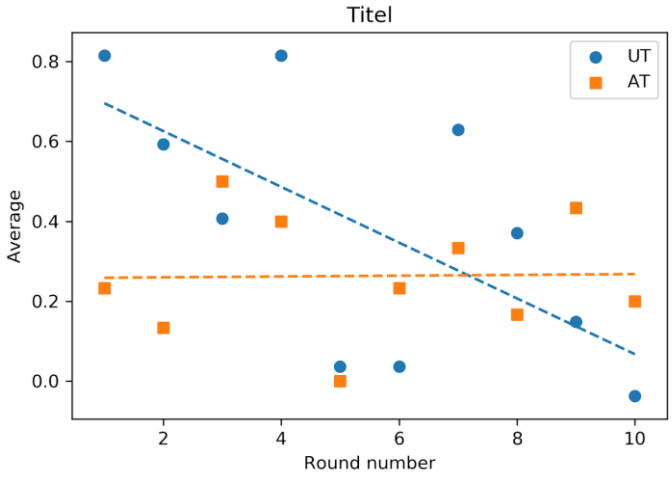
The data we obtained was analyzed with R-3.6.1 and plotted in Python 3.7. Both scripts can be found in the appendix.

A. Quantitative results

The experiment was conducted to find out if there is a significant difference between user satisfaction levels about the responses given by AT and the responses given by UT, and whether there is a significant difference in which player the user chose to play another round with in the future. To find out if the differences between these variables were significant, a paired-samples t-test was conducted to compare the average satisfaction about AT and UT measured over all the rounds. No significant difference was found in the satisfaction scores for AT (M=0.253, SD=1.28) and UT (M=0.29, SD=1.64); t(29)= -0.187, p = 0.853.

Another paired-samples t-test was conducted to measure whether the participants chose AT significantly more often than UT by the end of the experiment. A significant difference was found between how many participants chose AT (M=0.7, SD=0.47) and UT (M=0.3, SD=0.47); t(29)= 2.53, p = 0.0258. There were no significant effects of any of the control variables on the variables of interest, except a strong positive correlation between how satisfied the participants were with the amount of messages given by UT and the likelihood that they chose UT, which was expected: if people are not frustrated but satisfied by UT's many apologies, they are more likely to choose UT to play with in the future.

a) Satisfaction response AT and UT per round: As expected, the t-test of the satisfaction levels after each round came back with no significant results. Since we were expecting a turning point, starting out with users not being frustrated by UT's answers but getting less satisfied after a certain amount, the numbers should even out; unless the satisfaction levels of AT would have shown the same results in the opposite direction, a statistical difference between the two means was not necessarily to be expected. More interesting for this variable is the plotting of the satisfaction throughout the rounds to see how satisfied the users were about AT and UT in the beginning of the experiment and whether this changed.



Plot 1. Progression of average satisfaction levels about the responses given by AT and UT (least squares fit). Higher numbers signify more satisfaction and less frustration while lower numbers signify more frustration and less satisfaction.

To find out if the responses differed significantly over the rounds, a one-way ANOVA test was conducted. With an F-value of 38.19 ad a P-value of 1.13e-06, we can state that there is a statistical difference in the progression of the two variables.

As visible in plot 1, the average frustration levels about the responses of teammate AT largely stay the same. We can see some fluctuations in the progression as the frustration levels go up and down, but there is no clear trend. For UT, we do see a trend: users start off feeling more satisfied with UT's responses than with AT's responses and gradually start getting less satisfied with UT's responses over the course of the experiment, while the satisfaction about AT's responses stay the same.

b) Turning point: One thing that stands out about these results is that the turning point that we expected is not obvious: the trend moves downward from the beginning. We could, however, say that there is a turning point where the users move from neutral (which is the way they are feeling about AT the whole time) to negative, which is around round 7. It is more important however to note that the users apparently start getting less satisfied with the unconditional responses from the second time they appear.

c) Satisfaction performance AT and UT per round: Our measuring of the satisfaction about the performance of AT and UT after every round yielded no significant results. The plot can be found in the appendix.

d) Teammate A and B overall: With a p-value of 0.0258, participants chose AT to play another round with in the future significantly more often than UT, indicating a higher level of satisfaction towards AT overall.

B. Qualitative results

At the end of the experiment, we asked each user why they had chosen teammate AT or UT to play with again in the future and got a diverse collection of responses.

a) Users who chose AT: Many users were frustrated by the number of unjustified apologies they received from UT, causing them to be annoyed with UT and therefore choosing AT. There was however also a reasonable amount of users that chose AT because they thought UT made more mistakes, even though this was not reflected in their answers about the satisfaction of the performance of AT and UT (see appendix). Since the computer teammates had the exact same level of skill, it seemed they were inclined to think UT was less able because it kept claiming responsibility for mistakes, even though it was visible to the users that AT and UT gave the same answers almost all of the time.

A few users who started with a positive attitude towards UT and got less satisfied with UT's responses, attested to being annoyed by UT's response because they felt they were contradicted in their own judgement of the situation.

b) Users who chose UT: Some users appreciated the constant feedback that was given by the unconditional apologies-teammate UT because they perceived the other as "too quiet". One user even noted that "I would rather have the wrong feedback than no feedback at all". Clearly there is a smaller group of users who prefer to receive more information from their teammate regardless.

VII. Limitations and Future Study

A. Diversity among participants

A limitation that could be interesting to tackle in further studies is diversity among the participants in age and nationality. In the group that we ended up studying we had a majority of participants that were female and between the ages of 20 and 30, and most of them were born in a more individualistic society as opposed to a more collective one. Although we had participants of seven different nationalities, the group was not divided equally. And even if the group was divided more equally, it would not be realistic to find effective evidence of within group differences in a group of only 30 participants. It might be interesting to devote separate research to error-handling and apologies among people of different ages and cultural backgrounds in the future.

B. Application in everyday systems: interval time

The crises that the user had to deal with in this study were artificial and constant, while in everyday situations crises appear in very different circumstances. The theory would have to be further tested in more natural error-handling circumstances. Important to note is that users were neutral to positive about the unconditional apologies in the beginning, indicating that if only one crisis pops up at a time, it doesn't matter that much whether the system takes responsibility and apologizes unconditionally or not. When the failures become more frequent is when the results of this study become useful. However, in future research we would have to take a look at how much interval time there can be between crises for the user frustration about the unconditional apologies to be reset (an hour? a day?) and whether there is a difference in this between different kinds of crises.

C. Application in everyday systems: ambiguity in responsibility

Another issue that needs to be mentioned when applying the results to real world systems is the ambiguity that realistically exists when it comes to responsibility (it is important to note here that the actual guilty party is always ambiguous and even irrelevant when talking about user experience: all that matters is who the guilty party is in the users eyes and how they should behave for optimal user experience). After all, if a user account is blocked because they typed the wrong password, this might be perceived by the user as their own mistake, but it could also be perceived as a fault of the system since it asked to change the password too often. In our experiment we purposely made the party that carried fault for the mistake ambiguous, to somewhat mimic a real-world crisis situation. However, it can still be easily inferred which party the real culprit is for the user than in real crisis situations. Therefore, in further research we could take a look at crisis situations that are even more ambiguous.

D. Generalizability: severity of consequences

A drawback of conducting an experiment on crisis situations in an artificial environment is that the results are not necessarily generalizable for all real-life situations. An issue for generalizability in this experiment would be by account of the severity of the crises. In this study, the users had the prospect of winning an extra prize if they managed to score well, but faced no dire consequences if they lost. The frustration pattern might be quite different when the user loses valued progress, information or time because of a failure they did or did not cause themselves; it has been shown that there is a significant relation between user frustration and the importance of the task [Caeparu 2004]. They also might be more or less inclined to accept the responsibility of the system and the apology. This would be an aspect to take into account when applying the results of the study to UX/HCI design and an area of interest for future study.

E. Generalizability: effect of the experimental design on results

Since the experiment was constructed in an artificial environment that cannot exactly mirror real life situations, it should be taken into account that certain aspects of the design have an effect on the result. What comes to mind is the fact that the user conducted the task with two computer teammates at the same time, while in real life situations there is not necessarily a "team" of three individuals present. The apologies of the UT might have an effect on the frustration levels about the response of the AT. If a real life situation arises where a user is handling an apology of a system separately, their responses might differ. For further research, it might be interesting to separate the responses.

VIII. Design implication

Although further research outside of the controlled environment of an experiment is necessary, we can make some general suggestions for improvements in system design that might lead to a less frustrating user experience based on this research. Apologeticness in a system is appreciated by users, but designers should take into account that if the mistakes happen frequently they should be selective with their apologies to avoid the impression that the apologies made by the system are empty or even cynical. They should also take into account that many apologies might make a system seem less capable or "insecure" to some of their users. Lastly, they should try to find out where their users put the responsibility of the mistakes that are made and design their systems accordingly.

IX. Conclusions

Finding out more about what causes satisfaction levels to rise and fall in users when mistakes are made can help system designers create a better user experience. In this study, we looked at the user satisfaction development when users were confronted with a teammate that apologized unconditionally and a teammate that apologized selectively in situations in which either the user or the system caused a failure that negatively impacted the user experience. We found that users were more satisfied with the responses by the unconditional apology-teammate the first time a mistake was made, but started to get more and more frustrated with every apology given, while the frustration about the adaptive apology teammate stayed neutral over all the rounds. Overall, when asked which teammate the users would want to play another round with in the future, a significantly larger amount chose the adaptive-apology teammate, indicating that in crisis situations where multiple mistakes are made, unconditional apologies yield less user satisfaction than apologies that are only given when the computer is considered at fault by the user.

X. Reflection

Before starting this thesis, my experience in doing experimental research was quite limited; except for two small projects, the main method of research in my Bachelors had been literary analysis. I set out to learn more about statistics, but more importantly about the process of designing not only a question, but an effective way to answer the question. Even though many different sources can be used, in literary analysis it seemed to me that the way of answering your question is laid out more clearly. In experimental research I felt that in finding the method to answering my research question the sky was really the limit, which was intimidating and confusing at first. Because where do you start? How do you know that the experiment you "randomly" came up with was a good way to start answering your question? With the support of my supervisors I started exploring different types of experiments that might work. In this we had to keep coming back to the research question at hand: does what I'm finding out now answer the research question? Or if the answer is always no, does the question need an upgrade? Countless times I was under the impression that I was finding out what I wanted to know, only to have to come back and realize I was finding answers to different questions entirely. It was this part of the process that was definitely the most challenging. Before I started, I was nervous about having to go up to people and ask them to be my test subjects, but this turned out to be a negligible part of a much larger and more complicated puzzle (and only took about two weeks). Overall, what I value most about the process of experimental research is the self-reliance, the feeling of finding answers building your own tools and shaping your own measures. That is not to say that we shouldn't build on the body of research that is already out there and use evidence, tools and measures that have been

established, but it seems to me that experimental research provides an extra layer of personal creativity to research.

REFERENCES

Aarts, H., Custers, R., & Wegner, D. M. (2005). On the inference of personal authorship: enhancing experienced agency by priming effect information. Consciousness and Cognition, 14(3), 439-458. doi:10.1016/j.concog.2004.11.001

Abeler, J., Calaki, J., Andree, K., & Basek, C. (2010). The power of apology. Economics letters, 107(2), 233-235. doi:10.1016/j.econlet.2010.01.033

Ames, D. L., & Fiske, S. T. (2013). Intentional harms are worse, even when theyr'e not. PsycEXTRA Dataset. doi:10.1037/e513702014-028

Bentley, J. M., Oostman, K. R., & Shah, S. F. (2017). Were sorry but its not our fault: organizational apologies in ambiguous crisis situations. Journal of Contingencies and Crisis Management, 26(1), 138-149. doi:10.1111/1468-5973.12169

Buhmann, J. M., Malik, J., & Perona, P. (1999). Image recognition: Visual grouping, recognition, and learning. Proceedings of the National Academy of Sciences, 96(25), 14203-14204. doi:10.1073/pnas.96.25.14203

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., & Shneiderman, B. (2004). Determining Causes and Severity of End-User Frustration. International Journal of human-Computer Interaction, 17(3), 333-356. doi:10.1207/s15327590ijhc1703_3 Chiles, B. W. (2015). Apologies and Apologizing. The International Encyclopedia of Interpersonal Communication, 1-9. doi:10.1002/9781118540190.wbeic224

Cowan, N. (2010). The Magical Mystery Four. Current Directions in Psychological Science, 19(1), 51-57. doi:10.1177/0963721409359277

Gardner, John. From Personal Life to Private Law. Oxford Scholarship Online, 2018, doi:10.1093/oso/9780198818755.001.0001.

Hewitt, J. P., & Tavuchis, N. (1992). Mea Culpa: A Sociology of Apology and Reconciliation. Contemporary Sociology, 21(4), 521. doi:10.2307/2075901

Lazare, Aaron (1995). Go Ahead, Say You're Sorry. Psychology Today, 1995.

Kito, M., Yuki, M., & Thomson, R. (2017). Relational mobility and close relationships: A socioecological approach to explain cross-cultural differences. Personal Relationships, 24(1), 114-130. doi:10.1111/pere.12174

Koeman, Lisa. "How many participants do researchers recruit? A look at 678 HCI/UX studies." Lisa Koeman, 17-06-2018, lisakoeman.nl/blog/how-many-participants-do-researchers-recruit-a-look-at-678-ux-hci-studies

Maddux, W. W., Kim, P. H., Okumura, T., & Brett, J. M. (2011). Cultural Differences in the Function and Meaning of Apologies. International Negotiation, 16(3), 405-425. doi:10.1163/157180611x592932

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 101(2), 343-352. doi:10.1037/0033-295x.101.2.343

Park, S. J., Macdonald, C. M., & Khoo, M. (2012). Do you care if a computer says sorry? Proceedings of the Designing Interactive Systems Conference on - DIS 12. doi:10.1145/2317956.2318067

Skarlicki, D. P., Folger, R., & Gee, J. (2004). When Social Accounts Backfire: The Exacerbating Effects of a Polite Message or an Apology on Reactions to an Unfair Outcome1. Journal of Applied Social Psychology, 34(2), 322-341. doi:10.1111/j.1559-1816.2004.tb02550.x

Teng, E. L., Taussig, M., Kempler, D., & Dick, M. B. (2010). Common Objects Memory Test. PsycTESTS Dataset. doi:10.1037/t34742-000

Wohl, M. J., Hornsey, M. J., & Philpot, C. R. (2011). A Critical Review of Official Public Apologies: Aims, Pitfalls, and a Staircase Model of Effectiveness. Social Issues and Policy Review, 5(1), 70-100. doi:10.1111/j.1751-2409.2011.01026.x

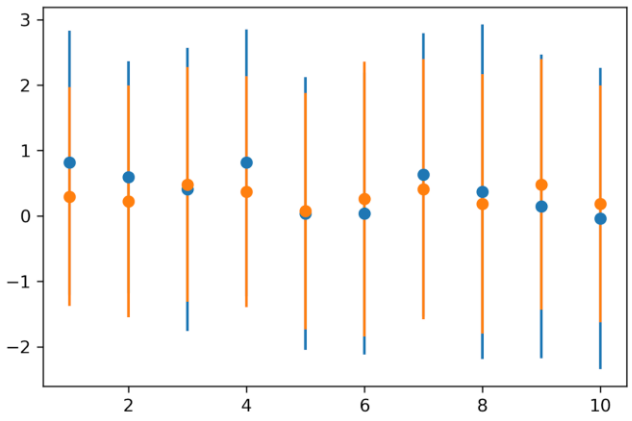
Zheng, X., Dijke, M. V., Leunissen, J. M., Giurge, L. M., & Cremer, D. D. (2016). When saying sorry may not help: Transgressor power moderates the effect of an apology on forgiveness in the workplace. Human Relations, 69(6), 1387-1418. doi:10.1177/0018726715611236Y.

Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

Young, M. The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

XI. Appendix

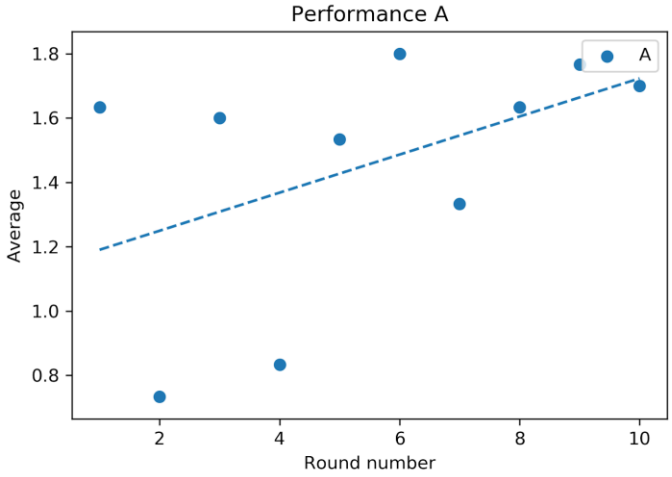
A. AT and UT progression with standard deviation



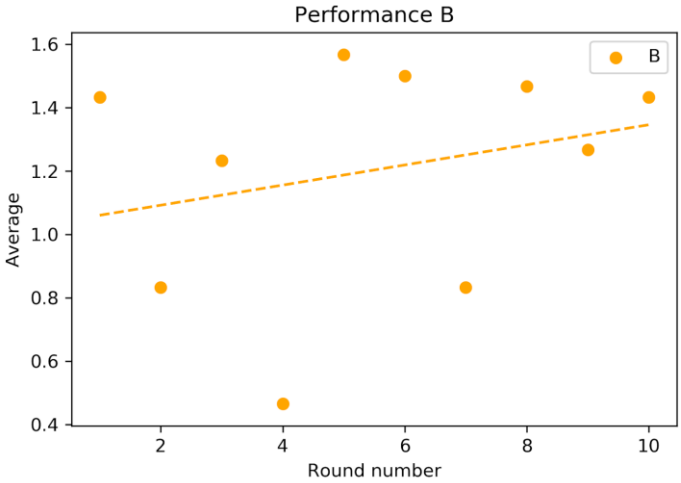
Plot 2. Progression of average frustration levels about the response given by A and B visualized with standard deviation

As we can see in this plot, the satisfaction levels from the users about teammate B's responses show a lot more fluctuations. It is clear that users are very divided on how frustrated they are by B's unconditional apologies, as users keep rating B's responses as both very satisfactory and very frustrating. Users are overall more neutral about A, and move from positive to negative about the responses by teammate B.

B. Perception of performance of AT and UT



Plot3. Progression of frustration about the performance ofteammate A over the course of the rounds.



Plot4. Progression of frustration about the performance of teammate B over the rounds.

The progression of the frustration levels about the performance of A and B are very similar, which was to be expected since A and B have the exact same level of skill at the game. As we can see in the graphs, the users rate their frustrations the same in the beginning (both around 1.4), and over the rounds the satisfaction over the performance of A increases by a few percent points, while the satisfaction over the performance of B decreases a little. The effect is not significant.

C. R code (t-tests and correlations)

https://drive.google.com/file/d/1QY65VNhb_dXckbv_7p3VVFwlZZXvkTCS/view?usp=sharing

D. Python code (scatterplots)

https://drive.google.com/file/d/1pLw9t3pnt2Bxmi4GGTHKnxpK5apJHfzX/view?usp=sharing

E. Consent form

https://drive.google.com/file/d/1AITFLZU3oHjurQ4ckebSv5NCm7PeMhMs/view?usp=sharing